SEMI-PARAMETRIC DENSITY ESTIMATION

by

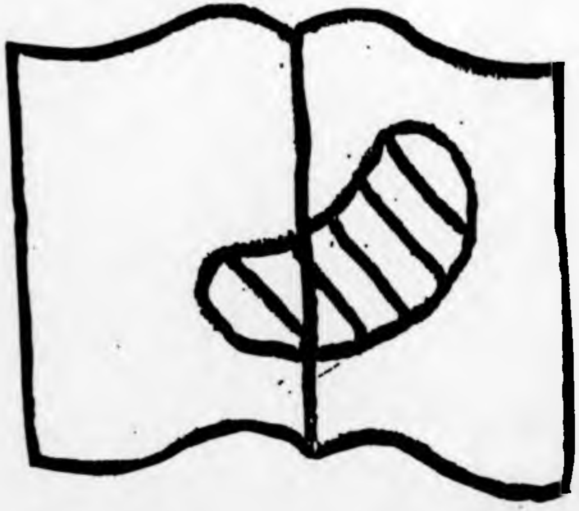
Christopher B. Stride, B.Sc.

This thesis is submitted for the degree of Doctor of Philosophy at the University of Warwick

Department of Statistics
University of Warwick
Coventry

September 1995

VARIABLE PRINT QUALITY



CONTENTS

PLOTS

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

ACKNOWLEDGEMENTS

Firstly I would like to thank my supervisor, Professor John Copas, for all his advice and encouragement over the past three years. Thanks also to all staff and students at Warwick University Statistics Department who have made the past six years very enjoyable, especially Lita and Liz (secretarial), Orlando, Nick and Crispin (for much needed computing help), Neil (rounds of golf) and Luis (mexican cooking)!

Finally to my parents and EPSRC whose financial support has enabled me to complete this work, and to God for giving me the ability to do it.

DECLARATION

I declare that this thesis is entirely the result of my own research during the past three years.

SUMMARY

The local likelihood method of Copas (1995a) allows for the incorporation into our parametric model of influence from data local to the point t at which we are estimating the true density function $g(t)$. This is is achieved through an analogy with censored data; we define the probability of a data point being considered observed, given that it has taken value x_i, as

$$w(x_i, t, h) = K\left(\frac{x_i - t}{h}\right)$$

where K is a scaled kernel function with smoothing parameter h. This leads to a likelihood function which gives more weight to observations close to t, hence the term 'local likelihood'.

After constructing this local likelihood function and maximising it at t, the resulting density estimate $f(t, \hat{\theta}_t)$ can be described as semi-parametric in terms of its limits with respect to h. As $h \to \infty$, it approximates a standard parametric fit $f(t, \hat{\theta})$ where as when h decreases towards 0, it approximates

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

the non-parametric kernel density estimate.

My thesis develops this idea, initially proving its asymptotic superiority over the standard parametric estimate under certain conditions.

We then consider the improvements possible by making smoothing parameter h a function of t, enabling our semi-parametric estimate to vary from approximating $\hat{g}(t)$ in regions of high density to $f(t, \hat{\theta})$ in regions where we believe the true density to be low. Our improvement in accuracy is demonstrated in both simulated and real data examples, and the limits with respect to h and the new adaption parameter α are examined. Methods for choosing h and α are given and evaluated, along with a procedure for incorporating prior belief about the true form of the density into these choices.

Further practical examples illustrate the effectiveness of these ideas when applied to a wide range of data sets.

NOTATION USED IN THIS THESIS

The following notation appears regularly throughout this thesis. This guide is neither exhaustive or essential, but may save the reader the trouble of referring back to preceding chapters when working through equations.

g The true distribution whose density function we are attempting to estimate.

$X = x_1, ..., x_n$ A random sample of n observations from distribution g.

t Target point t at which we are estimating the density function of the true distribution.

$g(t)$ The probability density function defining true distribution g.

n The size of our sample of random observations.

$f(t, \theta)$ Probability density function defining the parametric family f which **we believe** to be the distribution which produced data set X.

$\hat{\theta}$ Maximum Likelihood Estimate (MLE) of parameter θ.

$\hat{\theta}_t$ Maximum Local Likelihood Estimate (MLLE) of parameter θ at t.

$K(u)$ Scaled kernel function performing the weighting in our local likelihood function.

h Bandwidth or smoothing parameter used in the kernel function, constant with respect to t. 'Overall bandwidth' in adaptive semi parametric method.

h_t Local bandwidth; locally variable version of h.

α Adaption parameter which determines how much our local bandwidth varies from the overall bandwidth h.

$h_{t,\alpha}$ Local bandwidth formed using amount of adaption α

NOTATION USED IN THIS THESIS

The following notation appears regularly throughout this thesis. This guide is neither exhaustive or essential, but may save the reader the trouble of referring back to preceding chapters when working through equations.

g The true distribution whose density function we are attempting to estimate.

$X = x_1, ..., x_n$ A random sample of n observations from distribution g.

t Target point t at which we are estimating the density function of the true distribution.

$g(t)$ The probability density function defining true distribution g.

n The size of our sample of random observations.

$f(t, \theta)$ Probability density function defining the parametric family f which we believe to be the distribution which produced data set X.

$\hat{\theta}$ Maximum Likelihood Estimate (MLE) of parameter θ.

$\tilde{\theta}_t$ Maximum Local Likelihood Estimate (MLLE) of parameter θ at t.

$K(u)$ Scaled kernel function performing the weighting in our local likelihood function.

h Bandwidth or smoothing parameter used in the kernel function, constant with respect to t. 'Overall bandwidth' in adaptive semi-parametric method.

h_t Local bandwidth; locally variable version of h.

α Adaption parameter which determines how much our local bandwidth varies from the overall bandwidth h.

$h_{t,\alpha}$ Local bandwidth formed using amount of adaption α

$\hat{\theta}_{t,\alpha}$ MLLE of θ when the kernel function weighting the local likelihood function uses a local bandwidth $h_{t,\alpha}$ instead of h.

MSE Mean squared error.

MISE Mean integrated squared error.

ISE Integrated squared error.

1 Introduction and Literature Survey

1.1 Setting the scene

How exactly one should estimate the probability density function defining the true distribution g of a random quantity Z is a fundamental question in statistics. The various methods used for density estimation can be split into two groups. Parametric estimation assumes that the observed data $X = x_1, ..., x_n$ are from a member of a standard parametric family of distributions f, with density function $f(t, \theta)$ at our target point t at which we are attempting to estimate true density function $g(t)$. The p-dimensional parameter $\theta = (\theta_1, ..., \theta_p)$ is then estimated, for example by maximum likelihood, and an estimate $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_p)$ is obtained. We now have a parametric estimate $f(t, \hat{\theta})$. Non-parametric estimation differs from this in that no constraints are imposed as regards to the data coming from any particular parametric family. Instead the data are allowed to speak for themselves, the only assumption being that the distribution of the data has a 'true' probability density function $g(t)$, our non-parametric estimate of this being defined as $\hat{g}(t)$.

The aim of this thesis is to produce a technique incorporating the best aspects of parametric and non-parametric estimation, and eliminating their respective weaknesses as far as possible. Various methods have been proposed in the past which are semi-parametric in that they attempt to provide a link between $\hat{g}(t)$ and $f(t, \hat{\theta})$. My work concentrates on the development, improvement and theoretical justification of a particular technique built on the concept of **local likelihood**. This method was first applied in a density estimation setting by Copas (1995a), though local weighting of observations has been established in the field of regression analysis for some time.

1.2 Local likelihood and regression

Local fitting is a standard procedure in estimating the dependence of a response variable Y on a predictive variable X. Assume we have a sample $(X_1, Y_1), ..., (X_n, Y_n)$. We are searching for a regression function

$$m(X) = E(Y|X),$$

the best mean squared error predictor of Y given X, and our response shall then be modelled as

$$Y_i = m(X_i) + \epsilon_i$$

where the mean of the ϵ_i's is 0 for each i.

Cleveland (1979) and Chambers (1983) both give examples of Scatterplot Smoothers. For a scatterplot (X_i, Y_i), $i = 1, ..., n$, the fitted value at x is the value of a polynomial fit to the data using weighted least squares. The weight of a particular point is dependent upon the distance between the X variables only, such that it will be large for (X_i, Y_i) if X_i is close to x, and small if not. A common choice of weight function is

$$w(x, X_i, h) = K\left(\frac{x - X_i}{h}\right) \tag{1}$$

where $K(u)$ is a scaled non-negative kernel function which satisfies the following conditions:

(i) It is symmetrical about 0.

(ii) It is a decreasing function of $|u|$.

(iii) It has a maximum of 1 at $u = 0$.

For example, if we are estimating response y at predictor value $x = x_0$, we fit a straight line

$$y = \hat{a}_0 + \hat{b}_0 x \tag{2}$$

where $\hat{a}_0$ and $\hat{b}_0$ are the values of coefficients a and b which minimise

$$\sum_{i=1}^{n}(Y_i - a - bX_i)^2 K\left(\frac{x - X_i}{h}\right)$$

at $x = x_0$. Smoothing parameter h, often called the **bandwidth**, controls the degree of local weighting; when it is large this method approximates ordinary least squares. Our regression estimate at x_0 is simply the height of the fitted line (2) at x_0. At a different point x_1 we repeat the procedure, but with the weighting now dependent on the distances of the X_i's from x_1. Plot 1a, adapted from Wand and Jones (1995), Fig 1.2, illustrates this method clearly.

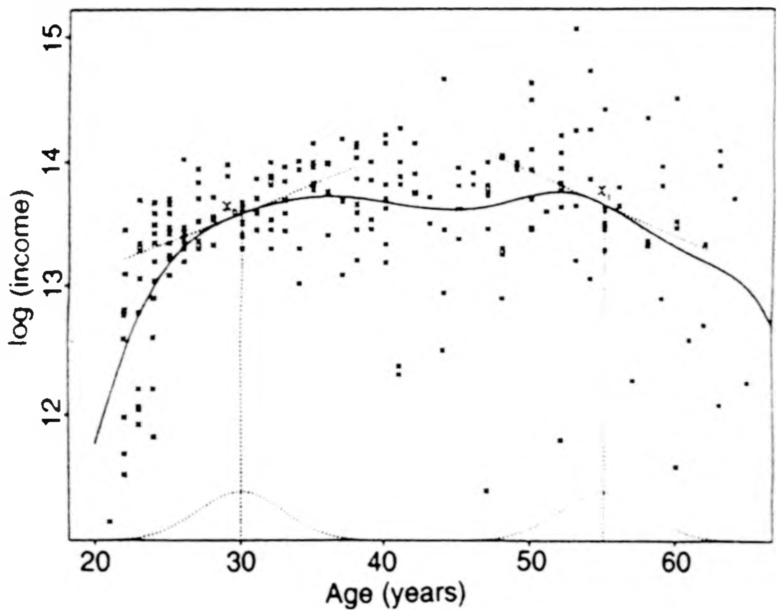
The result is a non-linear fit to the data, with the extent of the departure from linearity depending on the size of h. This method is also known as the locally weighted running lines smoother or LOWESS. Cleveland (1979) gives a more robust iterative fitting method.

Scatterplot smoothing can be seen as semi-parametric in that we are assuming a model for Y given X at every point. In general the literature describes this as non-parametric regression.

Tibshirani and Hastie (1987) make the extension to likelihood based regression models. Likelihood based methods assume that $m(X_i) = m(X_i, \beta)$ and then attempt to estimate parameter β. For example in the quadratic model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i \tag{3}$$

PLOT 1a: Scatterplot Smoother based on age/log(income) data



The solid curve is the estimate The dotted curves are the kernel weights and straight line fits at points x_0 and x_1

with

$$\epsilon_i \sim N(0, \sigma^2),$$

we estimate $\beta = (\beta_0, \beta_1, \beta_2)$ by $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$, the maximum likelihood estimate found by maximising log-likelihood function $L(\beta, X_i, Y_i)$ summed over the data. Our parametric estimate of response y given any predictor value x, found via likelihood, would thus be the curve

$$y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \epsilon_i.$$

Tibshirani and Hastie replace the function m of predictor value and parameter β in (3) by an unspecified smooth function $s(x)$. They then use 'local likelihood' to estimate $s(x)$ from the data. In this process the data is weighted with respect to x, the weight of an observed pair (X_i, Y_i) decreasing the further X_i is from x. We then find parameters $\hat{\beta}_x = (\hat{\beta}_{0,x}, \hat{\beta}_{1,x}, \hat{\beta}_{2,x})$ which maximise the 'local likelihood function'

$$\sum_{i=1}^{n} w(x, X_i, h) L(\beta_x, X_i, Y_i)$$

at predictor value x, where $w(x, X_i, h)$ controls the weighting. Our estimate of $s(x)$ is then defined as

$$\hat{s}(x) = \hat{\beta}_{0,x} + \hat{\beta}_{1,x} x + \hat{\beta}_{2,x} x^2.$$

In the Gaussian case, this local likelihood regression technique is equivalent to LOWESS, since maximising the local likelihood is equivalent to least squares minimisation. This procedure has a range of applications, such as in the proportional hazards model of Cox (1972). Hastie and Tibshirani (1990) discuss several drawbacks of this method, such as the difficulties of incorporating different smoothing methods and the cost of the estimation procedure. However they suggest that the method is worthy of further study.

Other work of interest in this field has been produced by Jianqing Fan, in particular Fan and Gibjels (1992) which explores the use of a variable bandwidth h_x. Explicit formulae for the optimal choice of this parameter are given.

1.3 Density estimation through local likelihood

In the words of Copas (1995a), there is a "crucial distinction" between the use of locally weighted likelihood methods in regression and in density estimation. In the former, the weight of an observation is a function of time or of the covariates which are fixed (i.e. we are interested in estimating response Y conditional upon fixed X), thus our weight function can be considered fixed. In the density estimation scenario, the weight needs to depend on the distance between the observed data X and the target point t at which we are estimating the true density. Because the weight function is a function of X and hence a random variable, we have to allow for the random nature of the weights in our likelihood function.

Copas uses an analogy with censored data to motivate a local likelihood function. As in parametric estimation we assume the data to be from a parametric family f. However, we consider a proportion of the data to have been observed and the rest to have been censored. Given that we are estimating at any target point t, the probability that any particular observation is taken as being observed, conditional on the value x_i that it takes arriving from the experiment, is given by a weight function

$$w(x_i, t, h) = K\left(\frac{x_i - t}{h}\right)$$

In other words, the weighting on each observation, which is dependent on its distance from t, is performed by a scaled kernel function $K(u)$ defined in equation (1). Smoothing parameter h controls the overall degree of censoring; given a fixed target point t, then if h is large with respect to $\max_{x_i} |x_i - t|$, the probability of being observed will be large for all observations, but when h

decreases, only the observations local to t are likely to be considered observed.

Considering our sample X to have been thinned by this artificial censoring process, we now construct our local likelihood function. The probability that an observation is observed **and** takes the value x_i is

$$p(\text{ith data point considered observed}|\text{ith data point takes value } x_i)p(\text{a data point takes value } x_i)$$

$$= w(x_i, t, h)f(x_i, \theta).$$

Likewise the probability that a data point is censored and takes a value in the range of f can be written as

$$B(\theta) = \int_x (1 - w(x,t,h))f(x,\theta)dx = 1 - \int_x w(x,t,h)f(x,\theta)dx.$$

With I_i as the indicator random variable for the censoring status of observation x_i having been observed, then our standard log-likelihood is of the form

$$l(x_i, \theta) = I_i \log f(x_i, \theta) + I_i \log w(x_i, t, h) + (1 - I_i) \log B(\theta). \qquad (4)$$

We can omit the second term since it is constant with respect to model parameter θ. Then taking the expectation over the I_i's to get the 'average' log-likelihood obtained under our weighting mechanism, using the fact that

$$E(I_i|x_i) = w(x_i, t, h)$$

and summing over the data, we have the local likelihood function at t,

$$L_w(X, t, \theta, h) = \sum_{i=1}^{n} w(x_i, t, h) \log f(x_i, \theta) + (1 - w(x_i, t, h)) \log B(\theta). \quad (5)$$

For any target point t, we evaluate our maximum local likelihood estimate (MLLE) $\tilde{\theta}_t$ which maximises (5). Our density estimate is $f(t, \tilde{\theta}_t)$. Copas (1995a) gives further technical details, demonstrating that although the variance of the MLLE $\tilde{\theta}_t$ is greater than that of the standard maximum likelihood estimate (MLE) $\hat{\theta}$, $\tilde{\theta}_t$ is a consistent estimate of θ for all h.

This method is semi-parametric in terms of its limits with respect to h. Assuming n fixed, as $h \to \infty$ our local likelihood function converges towards the standard log-likelihood $L(X, \theta)$, which is maximised by MLE $\hat{\theta}$, and so our density estimate will approximate the standard parametric estimate.

It is important to note that this convergence is not always uniform for all t. That is to say, however large our value of h is, if the domain of f is unbounded then

$$t \to \infty \Rightarrow w(x_i, t, h) \to 0 \ \forall x_i,$$

and $L_w(X, t, \theta, h)$ will not approximate the log-likelihood. For practical use of this semi-parametric method this is not a problem, since we will either be interested in density estimation at a single point, or over a bounded interval which contains the data. In these cases $|x_i - t|$ will be bounded, and as $h \to \infty$ the aforementioned convergence is uniform. So if we want our local likelihood to approximate the log-likelihood for all t we just choose h to be very large with respect to $|x_i - t|$ for all the observations x_i.

However, later in this thesis I use this large h approximation to the local likelihood in a theoretical context, which involves integrating our semi-parametric estimate with respect to t over an unbounded interval. Despite the lack of uniform convergence of $L_w(X, t, \theta, h)$ to $L(X, \theta)$ as $h \to \infty$, it is possible to split the integrals concerned into three parts and demonstrate that the unbounded end sections converge to 0 as the probability of them containing observed data decreases. We can then use the bounded middle section integral as an approximation to the whole integral. This is explained in greater depth in section 2.3. Obviously if the domain of f is bounded anyway, then convergence is uniform and we do not have a problem.

Copas also demonstrates that as $h \to 0$, then $f(t, \bar{\theta}_t)$ converges to the ordinary non-parametric kernel estimate $\hat{g}(t)$ in the sense that as $h \to 0$, then

$$\frac{f(t, \bar{\theta}_t)}{\hat{g}(t)} \to 1$$

where

$$\hat{g}(t) = \sum_{i=1}^{n} \frac{1}{nhc} K\left(\frac{x_i - t}{h}\right). \tag{6}$$

In equation (6), which is our definition of $\hat{g}(t)$ for the remainder of this thesis, we have

$$c = \int_u K(u)du$$

as the normalising constant for kernel function K, $x_1, ..x_n$ as our sample and h as the smoothing parameter or bandwidth.

In the semi-parametric method, bandwidth h controls the relative influence of the imposed parametric model compared to that of the data. As it spans the continuum between 0 and ∞, so $f(t, \bar{\theta}_t)$ spans the continuum between $\hat{g}(t)$ and $f(t, \hat{\theta})$, though it should be noted that the non-parametric end of this is a 'moving target', since $\hat{g}(t)$ is itself dependent on h.

One drawback with this method is that $f(t, \bar{\theta}_t)$ does not necessarily integrate to 1. Obviously this condition is satisfied at the limits of h, and for the Normal distribution with large h it was proved by Copas (1995a) that

$$\int_t f(t, \bar{\theta}_t)dt = 1 + O(h^{-4}).$$

When the semi-parametric method has been used with a two-parameter family such as a Normal, Weibull or Gamma, the numerically calculated integral has been very close to one, whereas it has been noticeably less for a one parameter family such as the exponential. As Copas suggests, this may be because of the greater flexibility of the two parameter case in matching all values of g. **In all the examples in this thesis, the density estimate $f(t, \hat{\theta}_t)$ has not been normalised unless otherwise stated.**

Two further papers have been concerned with demonstrating the superiority of this semi-parametric method, using small and large h approximations to $f(t, \hat{\theta}_t)$ respectively. Copas (1995b) produces a small h approximation to the mean squared error of $f(t, \hat{\theta}_t)$ and demonstrates a gain in accuracy over the ordinary kernel estimate if f is modestly misspecified. Meanwhile, taking h to be large, Copas and Stride (1995) give a proof of the semi-parametric method's asymptotic improvement in accuracy over parametric estimation under certain restrictions on the form of f. This is contained in chapter 2 of this thesis.

There exist other so-called semi-parametric density estimators which have similar properties to the method described above. Olkin and Spiegelman (1987) build the bridge between $f(t, \hat{\theta})$ and $\hat{g}(t)$ in a simpler if somewhat ad-hoc manner, without the direct motivation through likelihood. They propose a density estimate of

$$\pi f(t,\hat{\theta}) + (1-\pi)\hat{g}(t),$$

a weighted sum of parametric and non-parametric estimates. This method also requires selection of a further parameter $\pi \in [0,1]$, possibly by cross-validation.

The semi-parametric method of Hjort and Jones (1994) bears a closer resemblance to that of Copas. At t they suggest choosing a value of θ to maximise their local likelihood function

$$L_n(t,\theta) = n^{-1}\sum_{i=1}^{n} K\left(\frac{t-x_i}{h}\right)\log f(x_i,\theta) - \int_x K\left(\frac{t-x}{h}\right) f(x,\theta)dx \qquad (7)$$

which is simpler than that of Copas (1995a), but lacking the direct motivation through weighted censoring of observations depending on their distance from our target point. The function defined in (7) has a useful property in that the parameter $\theta(t)$ which maximises its large n approximation at t also minimises a locally weighted version of the Kullback-Leibler distance between the true and the approximating parametric density at t. As $h \to \infty$, (7) converges to the standard log-likelihood minus 1, which is maximised by MLE $\hat{\theta}$. Equally when h decreases to moderate and small values so reducing the influence of the model, the resulting mainly non-parametric estimate has approximately the same variance as $\hat{g}(t)$ but a smaller bias term, resulting in a smaller mean squared error with respect to true distribution $g(t)$. Using the notation from (5), equation (7) can be rewritten as

$$L_n(t,\theta) = n^{-1}\sum_{i=1}^{n} w(x_i,t,h)\log f(x_i,\theta) - 1 + B(\theta).$$

Some rewarding conversations about this work with Dr Shinto Eguchi (Institute of Statistical Mathematics, Tokyo) led to his suggestion of the following local likelihood function, where we choose $\theta = \theta_t^T$ to maximise

$$L_T(t,\theta) = n^{-1}\sum_{i=1}^{n} w(x_i,t,h)\log f(x_i,\theta) - w(x_i,t,h)\log(1 - B(\theta)).$$

(In the special case where we are using a rectangular "neighbourhood-type" weight function, this method is equivalent to 'truncating' rather than censoring. The weight function determines the size of the interval around target point t containing the observations which we shall consider. We ignore all observations outside this region, thus truncating the data.)

This method gives its strongest results when h is large. As for the method of Copas, we can show that under certain distance measures the resulting semi-parametric estimate $f(t,\theta_t^T)$ offers an improvement in accuracy over parametric estimation (see chapter 2). However the result is stronger since in this case there are no restrictions on parametric family f. This truncation method loses out when h is decreased to a small value, since unlike the procedures of Copas or Hjort and Jones it does not possess the equivalence to, or the asymptotic improvement over non-parametric kernel estimation. If we believe the departure of $g(t)$ from $f(t,\theta)$ to be very slight, making a large value of h suitable, then this method could be effective, but it will perform poorly if h is small.

From now on in this thesis, the terms semi-parametric estimation and local likelihood will refer exclusively to $f(t,\bar{\theta}_t)$ and (5) respectively unless otherwise stated. **It should also be assumed that in all practical examples and simulations in which I have evaluated semi-parametric density estimates via maximising the local likelihood, I have taken the weighting function $K(u)$ to be the scaled Gaussian kernel function** $\exp(-\frac{u^2}{2})$. It has convenient continuity properties and makes it possible to find $B(\theta)$ analytically for many choices of f. Unless clearly stated otherwise, this assumption applies for the whole of this thesis.

1.4 Kernel density estimation and related methods

At the lower limit of smoothing parameter h, our semi-parametric method approximates to a non-parametric ordinary kernel density estimate $\hat{g}(t)$ with bandwidth h (6), first suggested by Parzen (1962). This itself has been subject to a large amount of research, much of which becomes relevant in the later chapters of this thesis. Silverman (1986) gives a good introduction to the subject as well as going into finer detail on several facets of it. Wand and Jones (1995) performs a similar role, with special attention given to the choice of the bandwidth h. As in the histogram, of which the kernel is simply a smoothed version, the choice of h will significantly effect the appearance of our density estimate. Parzen (1962) developed a plug-in formula for h based on minimising a small h approximation to the mean squared error of the estimate. This performs well when the true density is close to that of a Normal distribution, but otherwise tends to oversmooth, especially when estimating multimodal densities. Hall et al (1990) suggest an improvement involving a higher order asymptotic representation of the optimal bandwidth. Least Squares Cross-Validation, first applied in this context by Rudemo (1982), and Likelihood Cross-Validation (Duin (1976)) are also cited by Silverman as possible methods of selecting h. For a comparison of these methods see Sheather (1992), which demonstrates that there is no single best method for all samples.

Just as Fan suggests a variable h when using the kernel as a weight function in regression analysis, so the concept of kernel estimation with a variable h has also been explored. The principle of using a non-constant bandwidth is particularly appealing when true distribution g appears to be long-tailed. In these circumstances it makes sense to use a large bandwidth in areas of low density, smudging the few observations over a wide area and reducing the risk of noise around the datapoints. In the smaller regions of high density a smaller value of h would be more suitable to avoid oversmoothing and subsequent underestimation of the density around the mode.

Two methods have arisen, both of which can be described as 'adaptive' or 'variable' kernel methods. I shall use the definitions from the review paper of Jones (1990). This defines a varying kernel density estimate at t given data $X = x_1, ..., x_n$ as

$$\hat{g}_V(t) = \frac{1}{nc}\sum_{i=1}^{n}\frac{1}{h_{x_i}}K\left(\frac{t - x_i}{h_{x_i}}\right),$$

with the change from the ordinary kernel density estimate being that a different bandwidth is employed for each data point. Varying kernels were introduced by Briemen, Meisel and Purcell (1977), who suggested that the bandwidth would vary through a choice of $h_{x_i} = hg(x_i)^{-1}$ with h a constant of proportionality. Abramson (1982) proved that a choice of $h_{x_i} = hg(x_i)^{-\frac{1}{2}}$ reduced bias, while Silverman generalised the choice of bandwidth to

$$h_{x_i} = h\left(\frac{g(x_i)}{\exp(E_{x_i}(\log g(x_i)))}\right)^{-\alpha}$$

with new adaption parameter α chosen between 0 and 1. Notable further work in this area with respect to bias minimisation has been written by Hall and Marron (1988) and Hall (1990). Hall (1992) considers choosing our variable bandwidth with respect to minimising a weighted version of the integrated squared error of the resulting density estimate, and demonstrates the optimality of selection by weighted squared error cross-validation. Note that the formula for varying bandwidth h_{x_i} given above requires the estimation of a pilot estimate of the true density for further use. Brieman et al, Abramson and Silverman all indicate that the quality of our density estimate will be insensitive to the fine detail of the pilot estimate, for which Silverman suggests the use of ordinary kernel density estimation.

Local kernel estimation varies the bandwidth with location. At each target point t, our density estimate is

$$\hat{g}_L(t) = \frac{1}{nc}\sum_{i=1}^{n}\frac{1}{h_t}K\left(\frac{t-x_i}{h_t}\right).$$

This is equivalent to the ordinary kernel estimate at t with $h = h_t$. But unlike the varying kernel, which provides a global density estimate once we have placed the kernels of varying bandwidth over the n points, the local kernel estimate at t, which places kernels of identical bandwidth h_t over the data points, is not applicable at any other point than t. In fact our local kernel estimate over the range of t is made up of a continuum of individual ordinary kernel estimates with different bandwidths for each t. This means that $\hat{g}_L(t)$ does not necessarily integrate to 1, and is not a probability density function itself, unlike $\hat{g}(t)$ and $\hat{g}_V(t)$. The optimal choice of h_t will be the same as for the ordinary kernel estimate when considering estimation at a single point; we can use the same small h approximation to the MSE at t which tells us that our best choice is

$$h_t \propto \left(\frac{g(t)}{g''(t)}\right)^{\frac{1}{5}}.$$

See Jones (1990) for further details. Over the range of t this will run into difficulties when the second derivative of $g(t)$ is equal to zero. Schucany

(1989) attempts to sidestep this problem but his solution involves the awkward proposition of estimating the fourth and sixth derivatives of $g(t)$. For these reasons the varying kernel has a greater chance of being adopted as a regular method of density estimation; indeed Silverman (1986) demonstrates its worth in several examples. However Chapter 3 of this thesis involves local kernels in the weighting of the local likelihood function; a situation in which their lack of integrability to 1 is less important. A new method of bandwidth selection is proposed avoiding awkward derivative estimation, using a form similar to that for the varying kernel.

1.5 How this thesis aims to improve upon these methods

Very few distributions are perfectly modelled by any single parametric family. One answer is not to impose any parametric family on the data, and let it speak for itself by using non-parametric methods. Yet methods such as histograms are hardly practical for precise estimation of a density and even non-parametric kernel estimation is far from ideal in some situations. As well as lacking the convenient structure of parametric estimation, the problem of choosing a suitable bandwidth h often becomes a balancing act, where some roughness is allowed in the tails to enable a good estimate to be achieved in regions of high density.

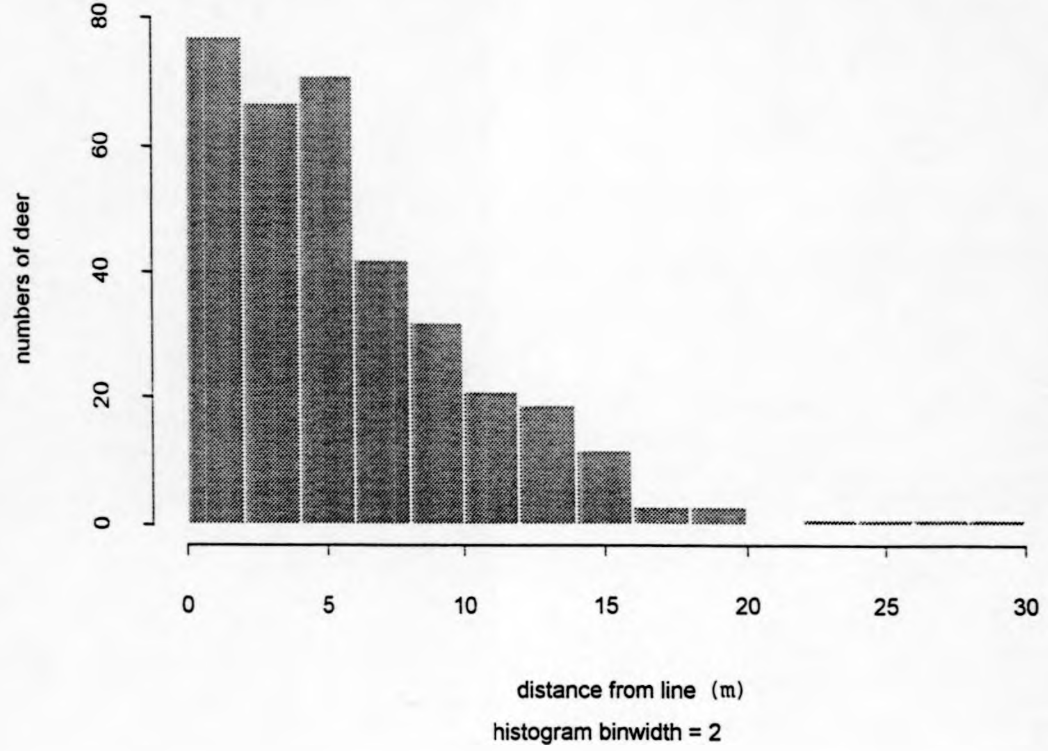
A motivation for semi-parametric methods of the sort reviewed in this chapter comes from cases where the true distribution is thought to resemble a parametric family but differs in shape and size in a few small areas. The following data set will be considered throughout this thesis since it is a good illustration of how both parametric and non-parametric methods can simultaneously fail.

The data set consists of line transect measurements of deer (Buckland (1992)). The unit of measurement of perpendicular distance from the line is metres. Line transect surveys are a technique used to measure 'species abundance', defined as the average number of specimens of the species per unit area. A straight line is drawn between two points and an observer walks along it, recording the perpendicular distances from the line of individual sightings of the relevant species. If L is the length of the line, then abundance can be estimated by

$$\hat{A} = \frac{ng(0)}{2L}$$

where n is the sample size and $g(t)$ is the probability density function of

PLOT 1b: Histogram of deer line transect data

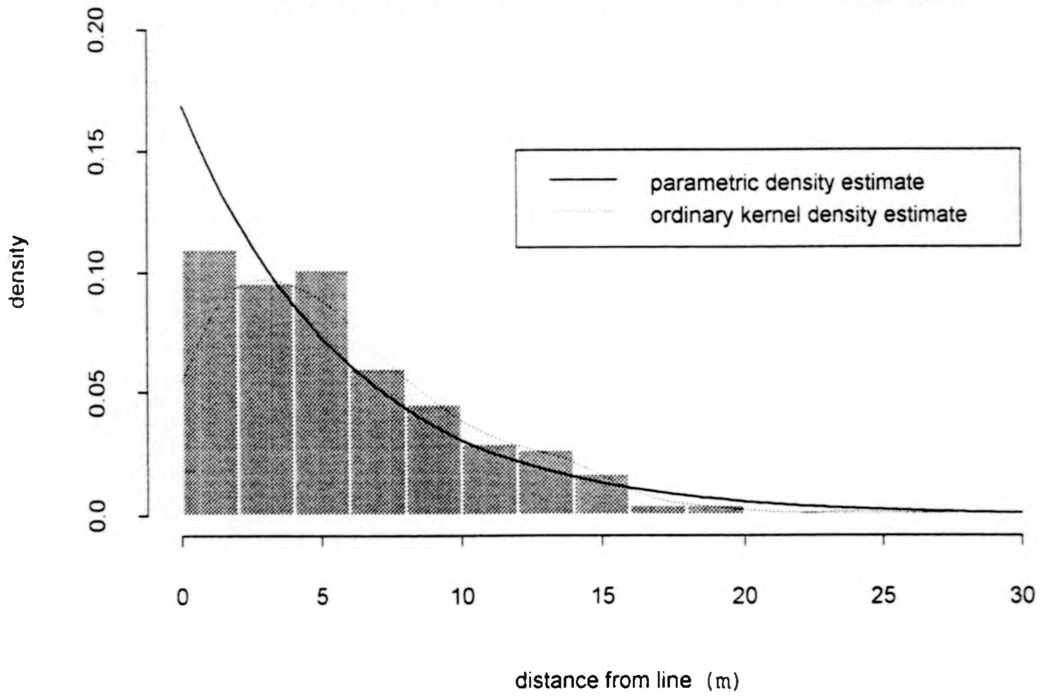


the observed distances, which itself must be estimated first. See section 2 of Copas (1995b) for a more detailed inspection of the formula for abundance.

Initially we consider producing a parametric estimate of $g(t)$. A histogram of the distances of the deer data (plot 1b) suggests an exponential fit is sensible, but overestimates in the region around $t = 0$. However, the alternative, a non-parametric kernel estimate, will struggle especially in this case, as much because of the shape and bounded nature of the distribution as the usual problem of choosing h to balance oversmoothing and undersmoothing. Thus thc retention of the structure of a parametric family which is defined on a limited range is convenient here; our density estimate cannot logically take any value other than 0 for $t < 0$ when we are considering distances from a line. Plot 1c illustrates just how poor our parametric and non-parametric estimates are around $t = 0$, the non-parametric method being an ordinary Gaussian kernel density estimate with bandwidth h handpicked to avoid noise in the tails yet minimise oversmoothing in regions of higher density. For a similar motivating example for semi-parametric methods see the remand times example from Copas (1995a), section 3. Further information on line transect sampling can be found in Burnham et al (1980).

Semi-parametric methods allow us to keep the neat parametric structure

PLOT 1c: Normalised histogram, parametric and kernel density estimates of deer line transect data



histogram binwidth = 2, kernel bandwidth (handpicked) = 1.5

but gives us flexibility to deal with local departures from the model. In particular the semi-parametric method derived from the local likelihood function (5) of Copas has the advantages of a clearly motivated structure, convergence to non-parametric and parametric estimates at the lower and upper limits of h, and asymptotic proof of its superiority over kernel estimation when $f(t, \theta)$ does not differ too much from $g(t)$. The first aim of this thesis is to demonstrate its superiority over parametric estimation under certain conditions (see chapter 2).

However the problem of choosing a suitable smoothing parameter h translates to semi-parametric estimation too, since we are using a kernel function for the weighting. Our 'balancing act' when selecting h is now between non-parametric and parametric estimates; we have to decide how much of the structure of the observed data we want to reveal compared to how close we want our estimate to stay to the standard parametric fit. This can pose problems in cases where we are certain that one region of the true density $g(t)$ will be akin to our parametric estimate $f(t, \hat{\theta})$, but another may differ noticeably. Chapter 3 of this thesis suggests a way of sidestepping this problem, by introducing a variable h allowing a more non-parametric estimate in some regions and a more parametric one in others. Having explained and developed a structure for this, chapters 4 and 5 proceed with the thorny question of how to select a value of h and a value of the new adaptive parameter α. Various ideas for plug-in values of these parameters based on minimising small h approximations to the mean squared error of our estimate are given. Chapter 6 suggests a revised approach, incorporating an index of prior belief concerning the true distribution. There are regular examples with real and simulated data sets throughout the thesis, with a couple of more detailed examples and conclusions contained in chapter 7.

2 The superiority of semi-parametric estimation when h is large

2.1 Introduction

The semi-parametric method of Copas (1995a), introduced and outlined in greater detail in chapter 1, combines the structure of parametric estimation through likelihood with the flexibility given by smoothing techniques that are normally associated with non-parametric procedures. Given a sample $x_1, ..., x_n$ assumed to be from model $f(x, \theta)$, the usual log-likelihood function

$$L(\theta) = \sum_{i=1}^{n} \log f(x_i, \theta)$$

is replaced by a local likelihood function (chapter 1, equation (5)) which gives extra weighting to the observations in the region of our target point t, the location at which we are estimating the true density. If all the weight in the analysis is placed on observations close to t, then the model is less influential, and we will obtain a mainly non-parametric density estimate. As the weighting is reduced, the model becomes more influential, with the evenly weighted case being equivalent to the full parametric fit produced by choosing $\theta = \hat{\theta}$ which maximises $L(\theta)$.

The bandwidth h smoothing kernel function $K(u)$, which performs the weighting within the local likelihood function, has overall control of this weighting. If h is large then the weighting will be even throughout the observations, but if h is very small then all the weight is placed on observations close to our target point.

Useful approximations to the equations given in chapter 1 can be found for both large and small values of h. For the latter it can be proved that provided n is sufficiently large, the asymptotic mean squared error (MSE) of the semi-parametric estimate must be smaller than that of the parametric fit whenever the model is misspecified. Details of the argument proposing this can be found in Copas (1995b), section 3.

We can consider the size of h to be relative to the scale of the data. For example, if h is large we interpret this as v being large, where $h = sv$, with $s = \hat{\sigma}$ denoting one sample standard deviation, and v constant. The approximations to be examined in more detail in section 2.2 rely on large h in terms of assuming that $(x-t)h^{-1}$ is small. Given a fixed value of h for all t, as $t \to \infty$ we will find that this assumption doesn't hold. However large h is, we can always choose t such that $(x-t)h^{-1}$ is not small. This dilemma can only be resolved if h is chosen relative to the distance between the target point t

and the data point furthest from t. Since this chapter aims to substantiate the advantage of the semi-parametric method with large **fixed** h over a range of points, which involves integrating over t from infinity to minus infinity, we need to use another approximation as $|t|$ gets very large. This is defined at the start of section 2.2.

Copas (1995a) derives several formulae based on fixed large h approximations of the weight function which determines the extent of the censoring of observations. For example, we can replace the kernel function $K((x-t)h^{-1})$ by its large h expansion, explained in more detail in section 2.2, producing the following equation,

$$L_w(t, X, \theta, h) =$$

$$\sum_{i=1}^{n} \log f(x_i, \theta) - \frac{1}{2h^2}(x_i - t)^2 \left(\log f(x_i, \theta) - \log({\sigma_\theta}^2 + (\mu_\theta - t)^2)\right) + O(h^{-4}). \tag{1}$$

Differentiating and omitting all terms of order greater than h^{-2} we find that

$$\frac{d}{d\theta} L_w(t, X, \theta, h) \simeq \frac{d}{d\theta} L(\theta) + \frac{1}{2} b\delta^2 nT \tag{2}$$

where $b = K''(0) \leq 0$, $\delta = h^{-1}$, n is the sample size and

$$T = n^{-1} \sum_{i=1}^{n} \left((x_i - t)^2 \left(\frac{d}{d\theta} \log f(x_i, \theta) - \frac{d}{d\theta} \log({\sigma_\theta}^2 + (t - \mu_\theta)^2) \right) \right) \tag{3}$$

with μ_θ and $\sigma_\theta{}^2$ being the mean and variance under density $f(x,\theta)$.

Under the assumption that model $f(x,\theta)$ is correct, T has mean 0, but if model f does not fit well in the neighbourhood of t, then $E(T) \neq 0$. Thus T can be used as a test of local fit, with the null hypothesis being that its true value is 0. Copas (1995a) pursues this further.

By considering these approximations we are seeing how much our MLE $\hat{\theta}$ is modified by a small amount of 'local influence' from the data. If $T \approx 0$, then at any target point t we would expect little departure of $\tilde{\theta}_t$ from $\hat{\theta}$. However if the data local to t suggests that true distribution g differs substantially from our model f in this region, then $\tilde{\theta}_t$ will differ from $\hat{\theta}$. Intuitively we will expect this change to improve our estimate at t more often than not, with $f(t,\tilde{\theta}_t)$ being closer to $g(t)$ than $f(t,\hat{\theta})$ was, since it has given greater weight to the data in the neighbourhood of t.

In sections 2.2 and 2.3 large h approximations are extended and then applied in an attempt to quantify any general advantage of the semi-parametric method, if indeed any exists. When considering estimation at a specific target point t, using the semi-parametric method is not always preferable. As illustrated in section 2.5 example 1, there will always be at least one value of t at which the parametric estimate will equal the true distribution, a level of accuracy which the semi-parametric method may not be able to match and obviously not better. However, despite not being uniformly better over the whole range of t, the semi-parametric method can be shown to be superior under certain conditions when using some sensible loss functions to compare the performance of itself and the parametric method. Section 2.2 gives some results for when we assume both n and h to be large.

2.2 Some approximations for large h

Several approximations for large h proposed in this section can be used to show the superiority of the semi-parametric method under certain conditions. Three loss functions will be employed to measure and compare the relative performance of parametric and semi-parametric estimation.

Given data $X = (x_1, ... x_n)$, the local likelihood function at target point t, $L_w(t, X, \theta, h)$, is driven by our weight function

$$w(x_i, t, h) = K\left(\frac{x_i - t}{h}\right)$$

where bandwidth h controls the amount of smoothing. For all t we can find $\theta = \bar{\theta}_t$ which maximises $L_w(t, X, \theta, h)$. As h increases towards infinity, $\bar{\theta}_t$ will tend towards the parametric estimate $\hat{\theta}$. In this section we assume h to be large and thus $(\hat{\theta}_t - \hat{\theta})$ to be small for all t.

Assume that the data X is from an unknown true distribution with probability density function $g(t)$ at any target point t, with n large. If h is large enough such that $(x-t)h^{-1}$ is small, then

$$K\left(\frac{x-t}{h}\right) \simeq 1 + \frac{1}{2} b_1 \left(\frac{x-t}{h}\right)^2 + O(h^{-4}) \tag{4}$$

where $b_1 = K''(0)$.

We substitute in equation (4) for the weight function in the local likelihood

$$L_w(t, X, \theta, h) = \sum_{i=1}^{n} l_w(t, x_i, \theta, h) =$$

$$\sum_{i=1}^{n} w(x_i, t, h) \log f(x_i, \theta) + (1 - w(x_i, t, h)) \log \left(1 - \int_x w(x, t, \theta) f(x, \theta) dx\right).$$

to reach equation (1) after omitting an irrelevant term which is independent of θ. Note that we can now show, to the same order of approximation as (1), that

$$Var\left(\frac{d}{d\theta} l_w(t, x, \theta, h)\right) \simeq E\left(\frac{d}{d\theta} \log f(x, \theta)\right)^2 - b_1 h^{-2} J$$

where

$$J = \left({\sigma_\theta}^2 + (t - \mu_\theta)^2\right)^{-1} \left(\frac{d}{d\theta}\left({\sigma_\theta}^2 + (t - \mu_\theta)^2\right)\right)^2 -$$
$$E\left((x-t)\frac{d}{d\theta} \log f(x, \theta)\right)^2,$$

and hence, using the standard variance approximation

$$Var(\tilde{\theta}_t) = n^{-1}\left(E\left(\frac{d^2}{d\theta^2}l_w(t,x,\theta,h)\right)\right)^{-2} Var\left(\frac{d}{d\theta}l_w(t,x,\theta,h)\right),$$

that

$$Var(\tilde{\theta}_t) \simeq \left(-nE\left(\frac{d^2}{d\theta^2}\log f(x,\theta)\right)\right)^{-1} + O(h^{-4}).$$

Thus when h is large, the variance of $\tilde{\theta}_t$ will be close to that of θ.

Continuing in a similar vein, we can also can also derive equation (2) from (1), and from this the following approximation,

$$(\tilde{\theta}_t - \hat{\theta}) \simeq -\frac{1}{2}\delta^2\left(E_f\left(\frac{d^2}{d\theta^2}\log f(X,\theta)\right)\right)^{-1} T, \tag{5}$$

where $\delta = h^{-1}$ and T is defined in section 2.1, equation (3).

$(\theta - \hat{\theta})$ is of order $n^{-\frac{1}{2}}$ so as $n \to \infty$ then $\hat{\theta} \to \theta$. With b_1 as above, we define

$$I(\theta)^{-1} = b_1\left(E_f\left(\frac{d^2}{d\theta^2}\log f(x,\theta)\right)\right)^{-1}$$

which is positive semi-definite. For example, if f is a Normal density with variance σ^2 and $K(u)$ is the Gaussian kernel, then

$$I(\theta)^{-1} = \sigma^2\begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}. \tag{6}$$

Thus, using (3) and (5) we find

$$E_g(\tilde{\theta}_t - \hat{\theta}) \simeq$$

$$-\frac{1}{2}\delta^2 I(\theta)^{-1} \int_x (t-x)^2 \left(\frac{d}{d\theta} \log f(x,\theta) - \frac{d}{d\theta} \log(E_f(t-x)^2) \right) g(x)dx \qquad (7)$$

which after expanding

$$= -\frac{1}{2} I(\theta)^{-1} \delta^2 \int_x (t-x)^2 \frac{d}{d\theta} \log f(x,\theta) g(x) dx$$

$$-\frac{d}{d\theta} \log E_f(t-x)^2 E_f(t-x)^2 c(t,\theta)$$

$$= -\frac{1}{2} I(\theta)^{-1} \delta^2 \int_x (t-x)^2 \frac{d}{d\theta} \log f(x,\theta)(g(x) - c(t,\theta) f(x,\theta)) dx \qquad (8)$$

where

$$c(t,\theta) = \frac{E_g(t-x)^2}{E_f(t-x)^2} = \frac{\int_x (t-x)^2 g(x) dx}{\int_x (t-x)^2 f(x,\theta) dx}.$$

We can find another useful approximation from the Taylor series of $f(t, \tilde{\theta}_t)$ expanded around $\hat{\theta}$, using the fact that $(\tilde{\theta}_t - \hat{\theta})$ is small for large h.

For all t we have

$$f(t, \tilde{\theta}_t) \simeq f(t, \hat{\theta}) + \left(\frac{d}{d\theta} f(t, \hat{\theta}) \right)^T (\tilde{\theta}_t - \hat{\theta}) + o(\delta^2). \qquad (9)$$

These approximations can now be used in evaluating the performance of the semi-parametric method in estimating the true distribution g.

2.3 Comparing semi-parametric and parametric estimation for large h

2.3.1 Using the loss function L_1

We shall now attempt to quantify the improvement gained by using the semi-parametric method, initially through loss function L_1. This is simply a weighted version of the integrated squared error (ISE), and when considering the accuracy of any parameter estimate $\theta = \theta^o$ over a range of target points t, it is defined as

$$L_1(f(t,\theta^o), g(t)) = \int_t \left(\frac{(f(t,\theta^o) - g(t))^2}{f(t,\theta)} \right) dt \tag{10}$$

where θ is the limit of Maximum Likelihood Estimate $\hat{\theta}$ as $n \to \infty$. Note that if we assume that $(\theta^o - \theta)$ is small for all t, then

$$L_1(f(t,\theta^o), g(t)) \simeq \int_t \left(\frac{(f(t,\theta^o) - g(t))^2}{f(t,\theta^o)} \right) dt, \tag{11}$$

the Kagan divergence between $f(t,\theta^o)$ and $g(t)$. See Clarke and Baron (1990) for further details of this function. It is analogous to the chi-squared goodness of fit statistic

$$U = \sum_r \frac{(F_r - E_f(F_r))^2}{E_f(F_r)},$$

which is distributed approximately ${\chi^2}_{k-1}$ if the data is divided up into approximately k groups. F_r is the number of observations in group r and $E_f(F_r)$

is the expected number in group r given that the data has true distribution f. The smaller U is, the better the fit of f to the data.

Thus the smaller the value of $L_1(f(t,\theta^\circ), g(t))$, the better our parameter estimate θ° is. With sample size n being very large, the variance of our estimate will be very small. Since the MISE can be written as a sum of the integrated variance and the integrated squared bias of $f(t,\theta^\circ)$, in the limiting case as $n \to \infty$ L_1 is measuring the weighted bias of our density estimate.

We can calculate the expectation of C_1, the difference between the performance under L_1 of the semi-parametric method (where $\theta^\circ = \tilde{\theta}_t$) and the parametric (where $\theta^\circ = \hat{\theta}$), defined as

$$
\begin{aligned}
C_1 &= L_1(f(t,\tilde{\theta}_t), g(t)) - L_1(f(t,\hat{\theta}), g(t)),\\
&= \int_t \left(\frac{(f(t,\tilde{\theta}_t) - g(t))^2 - (f(t,\hat{\theta}) - g(t))^2}{f(t,\theta)} \right) dt\\
&= \int_t \left(\frac{(f(t,\tilde{\theta}_t) - f(t,\hat{\theta}))(f(t,\tilde{\theta}_t) - g(t) + f(t,\hat{\theta}) - g(t))}{f(t,\theta)} \right) dt. \qquad (12)
\end{aligned}
$$

where the integral is evaluated over the domain of $f(t,\theta)$. To make further progress, we need to use the large h approximations to our semi-parametric method developed in section 2.2. These are really small $|x_i - t|h^{-1}$ approximations; it is possible to select a single fixed value of h to satisfy this criterion for all t in the domain of $f(t,\theta)$ if this domain is bounded. If it is unbounded we appear to have a problem, since for any fixed choice of h we can always find a large enough value of t within the domain such that $|x_i - t|h^{-1}$ is large too, contradicting the assumption underlying our approximations. The integral in equation (12) is calculated over the domain of $f(t,\theta)$; if this is unbounded we cannot choose h to validate the use of small $|x_i - t|h^{-1}$ approximations to $f(t,\tilde{\theta}_t)$ everywhere.

However, suppose, without loss of generality, that we examine the case where the domain of $f(t,\theta)$ is unbounded at both ends. Splitting up the integral gives

$$\begin{aligned}C_1 = &\int_{I1}^{I2} \left(\frac{(f(t,\tilde{\theta}_t) - f(t,\hat{\theta}))(f(t,\tilde{\theta}_t) - g(t) + f(t,\hat{\theta}) - g(t))}{f(t,\theta)} \right) dt \\ &+ \int_{-\infty}^{I1} \left(\frac{(f(t,\tilde{\theta}_t) - f(t,\hat{\theta}))(f(t,\tilde{\theta}_t) - g(t) + f(t,\hat{\theta}) - g(t))}{f(t,\theta)} \right) dt \\ &+ \int_{I2}^{\infty} \left(\frac{(f(t,\tilde{\theta}_t) - f(t,\hat{\theta}))(f(t,\tilde{\theta}_t) - g(t) + f(t,\hat{\theta}) - g(t))}{f(t,\theta)} \right) dt.\end{aligned}$$

As $|t| \to \infty$, we will pass the last data point and $K\left(\frac{t-x_i}{h}\right)$ will become small for all i and fixed h, so our semi-parametric estimate will approximate the kernel estimate $\hat{g}(t)$ defined in equation (6) of chapter 1. We will assume that $K(u)$ has been chosen to have tails at least as tight to zero as those of

$f(t,\theta)$, so that

$$|t| \to \infty \Rightarrow \frac{\hat{g}(t)}{f(t,\theta)} \to O(1).$$

When f is a Gamma, Weibull or Normal distribution, the popular Gaussian kernel function $K(u) = \exp(-\frac{1}{2}u^2)$ satisfies this restriction.

Consider values of $I1$ and $I2$ such that the integrals $\int_{-\infty}^{I1} f(t,\theta)dt$ and $\int_{I2}^{\infty} f(t,\theta)dt$ are both very small, and $|t - x_i|$ is very large for all t in the intervals $(-\infty, I1)$ and $(I2, \infty)$. Assuming throughout that n is large such that $\theta \simeq \hat{\theta}$, this implies that the second and third components of the sum of integrals given above can be approximated by

$$\text{Int } 1 = \int_{-\infty}^{I1} \left(\frac{\hat{g}(t)}{f(t,\theta)} - 1\right)(\hat{g}(t) - g(t) + f(t,\theta) - g(t))dt$$

and

$$\text{Int } 2 = \int_{I2}^{\infty} \left(\frac{\hat{g}(t)}{f(t,\theta)} - 1\right)(\hat{g}(t) - g(t) + f(t,\theta) - g(t))dt$$

respectively.

Then for sufficiently large $|I1|$ and $|I2|$,

$$0 < \frac{\hat{g}(t)}{f(t,\theta)} < v^*$$

for some constant v^*, so

$$
\begin{aligned}
|\mathrm{Int}\ 1| &= \left| \int_{-\infty}^{I1} \left(\frac{\hat{g}(t)}{f(t,\theta)} - 1 \right) (\hat{g}(t) - g(t) + f(t,\theta) - g(t)) dt \right| \\
&< \left| \int_{-\infty}^{I1} v^{*} (\hat{g}(t) - g(t) + f(t,\theta) - g(t)) dt \right| \to 0
\end{aligned}
$$

as $|I1| \to \infty$. The same applies for Int 2 in the positive limit case.

Hence, if we choose $h = v|I2 - I1|$, with v a large constant and $|I1|$ and $|I2|$ chosen very large as described above, then our large h approximations will be valid for

$$
\begin{aligned}
&\int_{I1}^{I2} \left(\frac{(f(t,\tilde{\theta}_t) - g(t))^2 - (f(t,\hat{\theta}) - g(t))^2}{f(t,\theta)} \right) dt \\
\simeq &\int_{-\infty}^{\infty} \left(\frac{(f(t,\tilde{\theta}_t) - f(t,\hat{\theta}))(f(t,\tilde{\theta}_t) - g(t) + f(t,\hat{\theta}) - g(t))}{f(t,\theta)} \right) dt = C_1. \quad (13)
\end{aligned}
$$

This solution to the problem of maintaining our approximations as $|t|$ increases towards infinity can be extended immediately to the case where we use the MISE as our loss function rather than L_1, which is explored in forthcoming subsection 2.3.3.

We can now return to equation (12) and use equation (9) to reach

$$
E(C_1) \simeq -2E_g \int_t (\tilde{\theta}_t - \hat{\theta})^T \frac{d}{d\theta} f(t,\hat{\theta}) \left(\frac{g(t) - f(t,\hat{\theta})}{f(t,\theta)} \right) dt. \qquad (14)
$$

The value $(\hat{\theta} - \theta)$ is of order $n^{-\frac{1}{2}}$, so that $\hat{\theta} \to \theta$ as n increases. Now assuming n is large causing $\hat{\theta} \simeq \theta$, and applying equations (7) and (8) to $E_g(\tilde{\theta}_t - \theta)$, we get

$$E(C_1) \simeq$$

$$\delta^2 \int_t \int_x (t-x)^2 \rho^T(t) I(\theta)^{-1} \rho(x)(g(t) - f(t,\theta))(g(x) - c(t,\theta) f(x,\theta)) dx dt, \quad (15)$$

where

$$\rho(x) = \frac{d}{d\theta} \log f(x,\theta).$$

Expanding this we find that if $c(t,\theta) = 1$, then $E(C_1) = -2\delta^2 z^T I(\theta)^{-1} z \leq 0$, where z is a vector, implying that the semi-parametric method is at least as accurate as the parametric under this loss function if $E_f(t-x)^2 = E_g(t-x)^2$ This condition holds in cases where f matches the mean and variance of any distribution to which it is fitted. For example, this occurs **if f is a Normal distribution with mean μ and variance σ^2.** In this case $I(\theta)^{-1}$ is as given in equation (6), and defining

$$\eta(x) = g(x) - f(x,\theta),$$

we have

$$E(C_1) =$$

$$\delta^2\sigma^2 \int_t\int_x (t-x)^2 \begin{pmatrix} \frac{d}{d\mu} log f(x,\theta)\eta(x) \\ \frac{d}{d\sigma} log f(x,\theta)\eta(x) \end{pmatrix}^T \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \frac{d}{d\mu} log f(t,\theta)\eta(t) \\ \frac{d}{d\sigma} log f(t,\theta)\eta(t) \end{pmatrix} dxdt$$

$$= -2\delta^2\sigma^2 \begin{pmatrix} \int_x x\frac{d}{d\mu} log f(x,\theta)\eta(x)dx \\ \int_x x\frac{d}{d\sigma} log f(x,\theta)\eta(x)dx \end{pmatrix}^T \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \int_t t\frac{d}{d\mu} log f(t,\theta)\eta(t)dt \\ \int_t t\frac{d}{d\sigma} log f(t,\theta)\eta(t)dt \end{pmatrix}$$

$$= -2\delta^2\sigma^2 \int_x x\frac{d}{d\mu} log f(x,\theta)\eta(x)dx \int_t t\frac{d}{d\mu} log f(t,\theta)\eta(t)dt$$

$$-\delta^2\sigma^2 \int_x x\frac{d}{d\sigma} log f(x,\theta)\eta(x)dx \int_t t\frac{d}{d\sigma} log f(t,\theta)\eta(t)dt$$

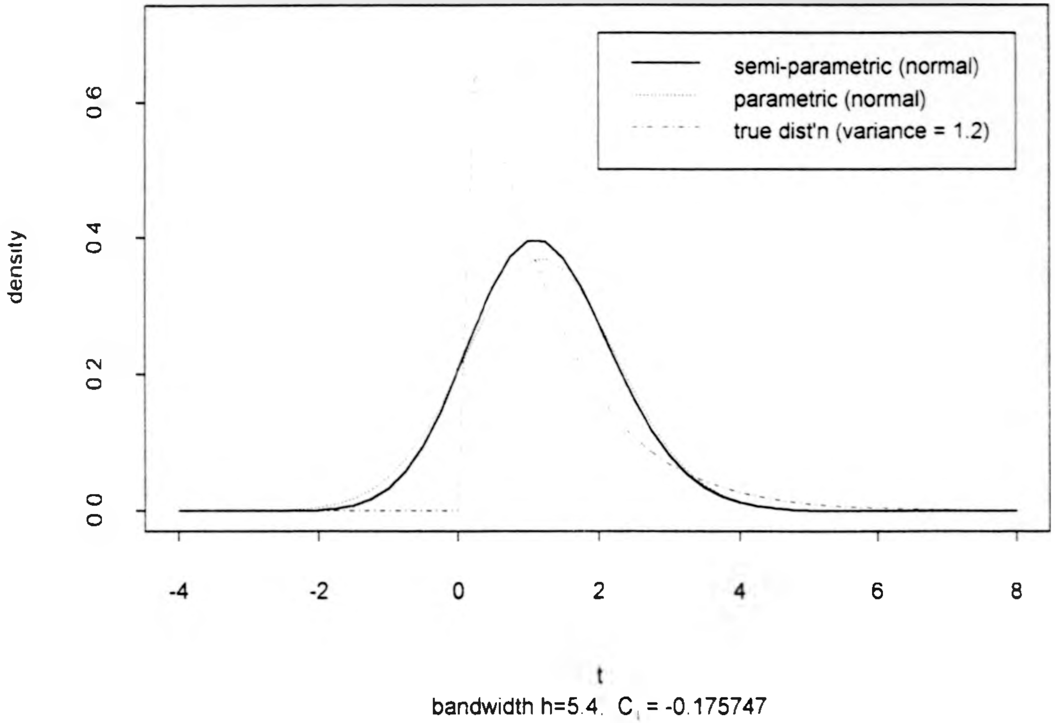
$$= -\delta^2\sigma^{-4}(F_3 - G_3)^2,$$

where F_3 and G_3 are the third moments of f and g respectively. As one would expect, a relationship exists between the gain in accuracy realised when choosing the semi-parametric method ahead of the parametric method, and the difference in the shapes of $f(t,\theta)$ and $g(t)$. The following graphs show this visually. In both cases I have taken a large random sample of 2000 points from a 'true' distribution g, chosen an 'incorrect' family parametric family f, and used a computer package to find $\hat{\theta}_t$ and $\hat{\theta}$ for my semi-parametric and parametric estimates $f(t,\hat{\theta}_t)$ and $f(t,\hat{\theta})$ respectively, over a large range of t's. Smoothing parameter (bandwidth) h is chosen large in both cases

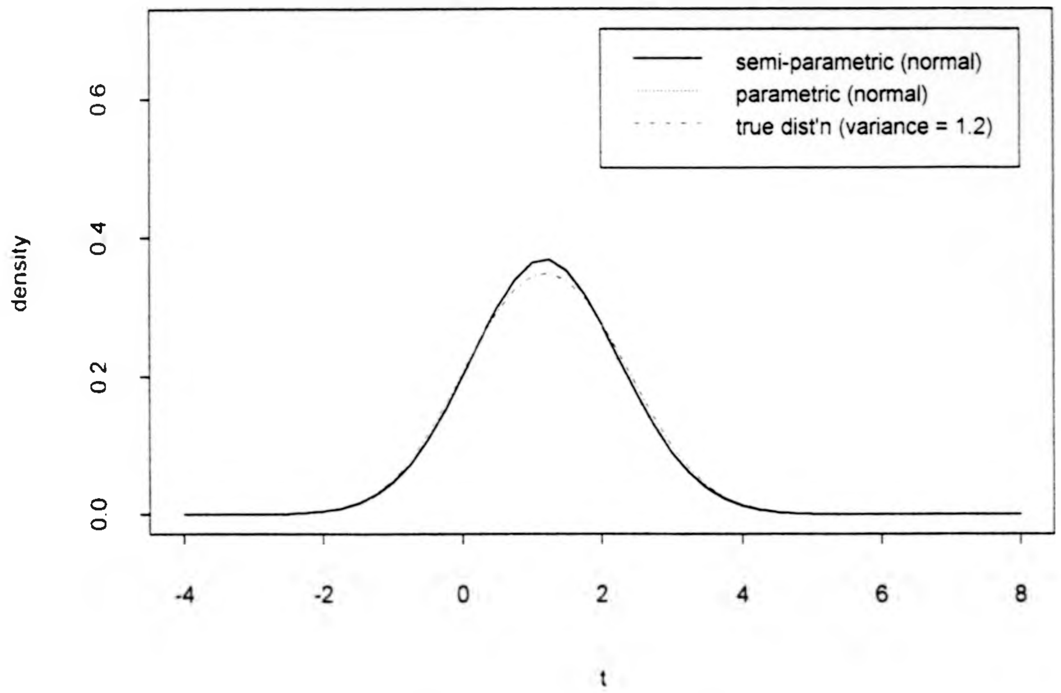
($h \simeq 5\hat{\sigma}$), complying with the assumptions of the earlier theory. The value of C_1 is estimated by numerical integration using my knowledge of g. To demonstrate that it is shape rather than variance which is related to the semi-parametric method's improvement, true distribution g has the same variance in both cases.

In Plot 2a the true distribution g is a Gamma[A,M] distribution, with parameters $A = 1$ and $M = \sqrt{1.2}$. It is highly skewed unlike f, a Normal distribution which we are trying to fit to the random sample from g. The application of a very small amount of local influence has made the semi-parametric method move slighly away from the parametric estimate and towards the true distribution in most regions. This is reflected in a much larger negative value of C_1 than in plot 2b where f is again a Normal distribution and g is a shifted Gamma with different shape parameters but with the same variance as in plot 2a. The shapes of $f(t, \hat{\theta})$ and $g(t)$ are similar with no substantial local departures of the true distribution from the model. Despite sampling variability, the semi-parametric and parametric estimates are virtually co-incident throughout the range of t.

PLOT 2a: Comparing parametric and semi-parametric estimates when shapes of g and f differ substantially



PLOT 2b: Comparing parametric and semi-parametric estimates when shapes of f and g are very similar



bandwidth h = 5.4. C_t = -4.882978e-006

2.3.2 Using the Kullback-Leibler Distance.

Suppose as our loss function L_2 we take the Kullback-Leibler distance between two distributions with density functions $g(t)$ and $f(t,\theta^\circ)$, namely

$$L_2(f(t,\theta^\circ),g(t)) = \int_t \log\left(\frac{g(t)}{f(t,\theta^\circ)}\right) g(t)dt. \tag{16}$$

Hall (1987) studies the properties of the Kullback-Leibler distance in measuring the accuracy of density estimates, though only in terms of non-parametric methods.

The Kullback-Leibler distance requires both density functions to integrate to 1. Semi-parametric estimate $f(t,\tilde{\theta}_t)$ does not necessarily possess this property; so we use the normalised semi-parametric estimate

$$\bar{f}(t,\tilde{\theta}_t) = \frac{f(t,\tilde{\theta}_t)}{\int_t f(t,\tilde{\theta}_t)}$$

instead.

Then another comparison of the relative accuracy of the semi-parametric and parametric methods can be achieved by calculating

$$E_g(C_2) = E_g(L_2(\bar{f}(t,\tilde{\theta}_t),g(t)) - L_2(f(t,\hat{\theta}),g(t))), \tag{17}$$

which can be written as

$$E_g\left(\int_t \log\left(\frac{f(t,\hat{\theta})}{\bar{f}(t,\tilde{\theta}_t)}\right) g(t)dt\right)$$

$$= -E_g\left(\int_t \log\left(\frac{f(t,\tilde{\theta}_t)}{f(t,\hat{\theta})}\right) g(t)dt - \log\int_t f(t,\tilde{\theta}_t)dt\right). \qquad (18)$$

To evaluate this we must integrate over a possibly unbounded interval of t, and use small $|(t-x_i)h^{-1}|$ approximations to $f(t,\tilde{\theta}_t)$. As in subsection 2.3.1, for any fixed h our 'small $|(t-x_i)h^{-1}|$' condition will be violated as $t \to \infty$. It is trivial to show that if two density functions are close everywhere then the Kagan Divergence between them, which in our example is approximated by L_1 for large n, is roughly twice the Kullback-Leibler distance between them. As $|t| \to \infty$, the differences $g(t) - f(t,\hat{\theta})$ and $g(t) - \bar{f}(t,\tilde{\theta}_t)$ between the true density and the parametric and semi-parametric estimates will both take very small values. Therefore we can approximate C_2 by $\frac{1}{2}C_1$ in the extreme tails, and use the theory of subsection 2.3.1 on page 37, which solves this limiting problem when we are evaluating loss function C_1.

Applying equation (9) and using the large h and n approximations

$$\log\left(1 + \frac{(\tilde{\theta}_t - \hat{\theta})\frac{d}{d\theta}f(t,\hat{\theta})}{f(t,\hat{\theta})}\right) \simeq (\tilde{\theta}_t - \theta)\frac{d}{d\theta}\log f(t,\theta)$$

and

$$\log\left(1+\int_t(\bar{\theta}_t-\hat{\theta})\frac{d}{d\theta}f(t,\hat{\theta})\right)\simeq\int_t(\bar{\theta}_t-\theta)\frac{d}{d\theta}\log f(t,\theta)f(t,\theta)dt,$$

we then have

$$E[C_2]\simeq-\int_t E_g(\bar{\theta}_t-\hat{\theta})^T\frac{d}{d\theta}\log f(t,\theta)(g(t)-f(t,\theta))dt, \tag{19}$$

which via equations (7) and (8) equals

$$\frac{1}{2}\delta^2\int_t\int_x(x-t)^2\rho(t)^T I(\theta)^{-1}\rho(x)(g(t)-f(t,\theta))(g(x)-c(t,\theta)f(x,\theta))dxdt. \tag{20}$$

If f satisfies $c(t,\theta)=1$ we once again have shown that the semi-parametric method is at least as accurate as the parametric. Note that in this case,

$$E(C_2)=\frac{1}{2}E(C_1) \tag{21}$$

indicating that again a relationship exists between the gain in accuracy when choosing the semi-parametric method ahead of the parametric method, and the difference in the shapes of $f(t,\theta)$ and $g(t)$. We would expect equation (20), having stated earlier that the Kagan Divergence between two functions is roughly twice the Kullback-Leibler distance between them if they are close everywhere in the region of interest.

2.3.3 Using the mean integrated squared error

Loss function L_1 given in equation (10) is effectively a weighted version of the more commonly used ISE between $g(t)$ and $f(t, \theta^\circ)$, defined as

$$ISE(f(t,\theta^\circ), g(t)) = \int_t (f(t,\theta^\circ) - g(t))^2 \, dt.$$

Define the mean integrated squared error (MISE) as

$$MISE(f(t,\theta^\circ), g(t)) = E_g\left(ISE(f(t,\theta^\circ), g(t))\right).$$

Attempting to use this to measure the expected difference in performance between the estimation methods, with the same conditions as before and using the same conditions as above, produced the following integral. In other words, if

$$C_3 = ISE(f(t,\tilde{\theta}_t), g(t)) - ISE(f(t,\hat{\theta}), g(t)),$$

then

$$E(C_3) = MISE(f(t,\hat{\theta}_t), g(t)) - MISE(f(t,\hat{\theta}), g(t)) \simeq$$

$$\frac{1}{2}\delta^2 \int_t \int_x (t-x)^2 \rho(t)^T I(\theta)^{-1} \rho(x)(g(t) - f(t,\theta)(g(x) - c(t,\theta) f(x,\theta)) f(t,\theta) dx dt \tag{22}$$

which I was unable to simplify algebraically for general $g(t)$, though I have shown this to be less than or equal to 0 for specific cases. For example if

$g(t)$ is the density function for an exponential distribution and $f(t,\theta)$ is the density function for a Normal distribution, then the integral can be calculated algebraically. If our true exponential distribution has parameter Λ, then

$$C_3 \simeq -0.303\delta^2\Lambda^{-1}$$

which implies that the semi-parametric method is better under these criteria and this measure. As suggested in Section 2.1, we can define $h = \delta^{-1}$ to be large when it is equal to v standard deviations of the true distribution g, where v is large. An exponential (Λ) distribution has variance Λ^{-2} so

$$C_3 = -0.303\delta^2\Lambda^{-1} = -0.303v^{-2}\Lambda.$$

As for loss functions L_1 and L_2, the larger the variance of the true distribution, the smaller the advantage we accrue by using the semi-parametric method instead of the parametric method. Alternatively this can be seen in terms of shape; the smaller the value of Λ (and thus our gain over the semi-parametric method), the flatter the density function will be, improving the Normal fit. Even then this is an admittedly impractical example, since it is unlikely that one would be attempting to fit a Normal distribution to data from an exponential distribution.

For the case where the true distribution g has a Gamma density, f is Normal distribution and all previously stated conditions hold, it is possible to obtain algebraically a formula for the integral in equation (24) in terms of the shape parameter M of the Gamma distribution. If $g \sim \Gamma[A, M]$, $f \sim N[\mu, \sigma^2]$, with smoothing parameter h and sample size n both large, then taking $A = 1$ without loss of generality we find that

$$E(C_3) \simeq \frac{1}{v^2 M} \frac{1}{2\sqrt{\pi} M} \left(\frac{-3}{2} \sqrt{M} \Phi(\sqrt{2M}) + \frac{\exp(-M)}{2\sqrt{\pi}} (-M(1+2M)) \right)$$

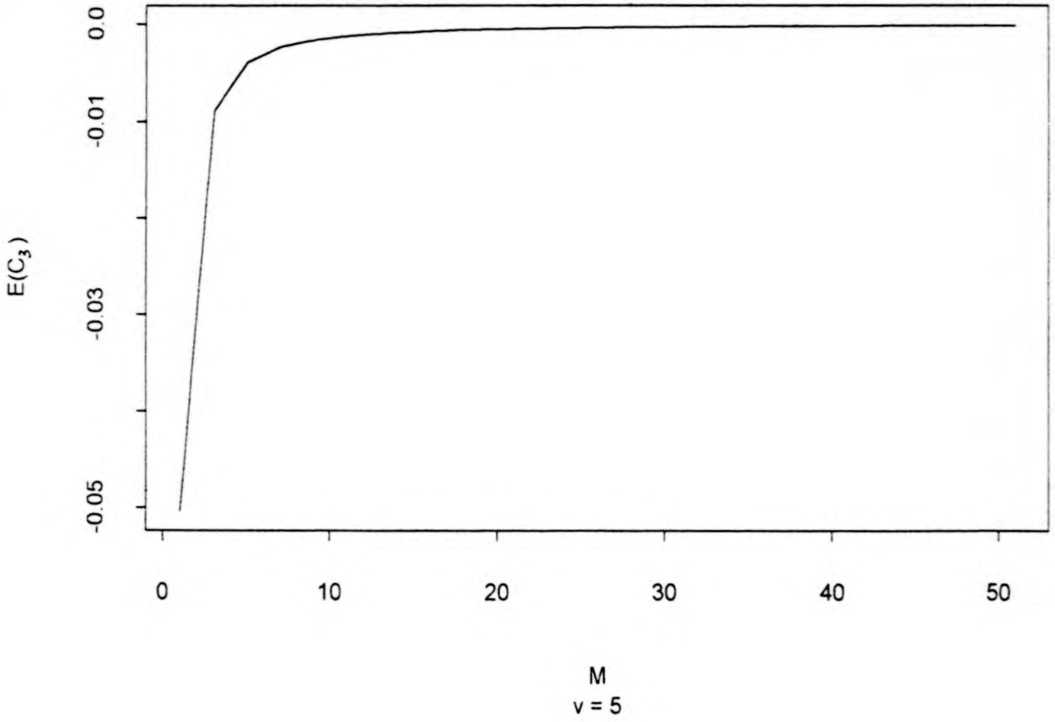
$$- \frac{\exp(\frac{-M}{2})(\sqrt{M})^{M-1}}{\Gamma(M+1)} \left((8M^2 + 4M)\sqrt{M}\Phi^*_M + (-8M^3 - 2M^2 + 3M)\Phi^*_{M-1} \right),$$

where $\Phi(t)$ is the normal c.d.f. and

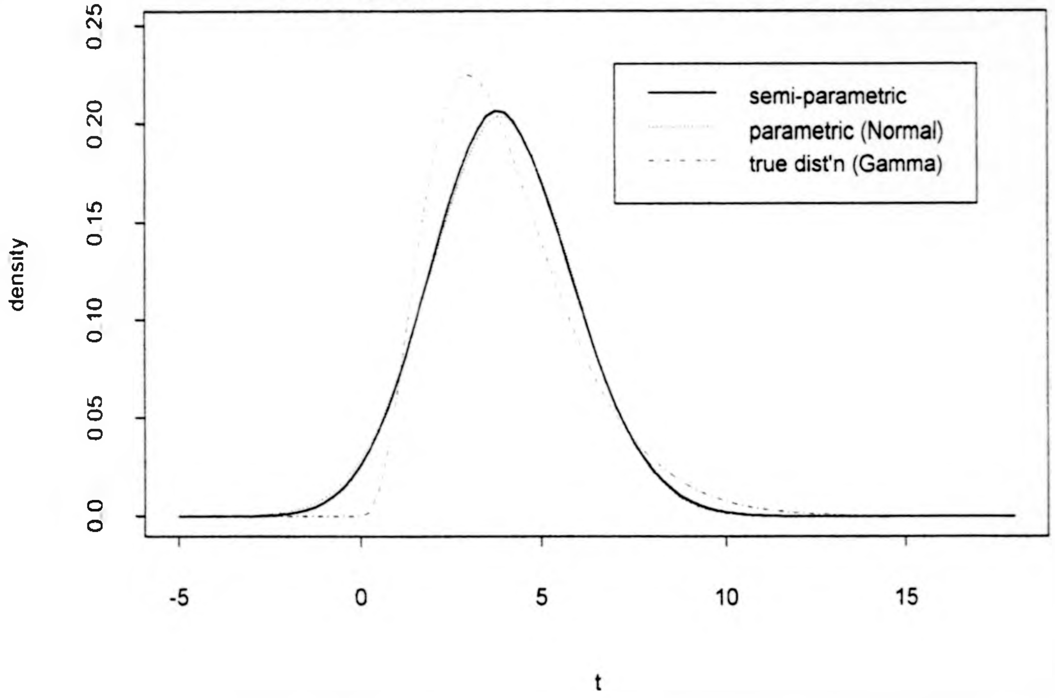
$$\Phi^*_M = \int_0^\infty u^M \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2}) du.$$

Plotting this formula against M we see that as M increases (and thus the variance of distribution g increases causing shape of $g(t)$ to become less skewed and closer to that of the Normal density function), then our gain in using the parametric method once again decreases (plot 2c). Note that for all M the semi-parametric method appears to outperform the parametric method. Thus both these results for the MISE support the comments made at the end of section 2.3.1. It should also be noted that both the above results are the MISE's calculated over the range $(0, \infty)$ i.e. only where $g(t)$ is defined. This gives more powerful results than those obtained by integrating over $(-\infty, \infty)$, since examples for the $g \sim \Gamma[A, M], f \sim N[\mu, \sigma^2]$ case (e.g. see plot 2d) indicate that in the left tail where $t < 0$ and $g(t) = 0$, $f(t, \hat{\theta}_t)$ is closer to 0 and thus a better estimate than $f(t, \hat{\theta})$. It suggests that integrating over $(-\infty, \infty)$ will produce results showing a much larger gain from using the semi-parametric method.

PLOT 2c: Comparing MISE's of parametric and semi-parametric density estimates when g~Gamma, f~Normal



PLOT 2d: Comparing parametric and semi-parametric density estimates under three different loss functions



bandwidth h=10, C_1= -0.03834373, C_2= -0.01904786 C_3= -0.001045033

2.4 Discussion

While we have no more proven examples, I would suggest that the loss of the semi-parametric method would be less than that of our parametric estimate under the MISE for at least as wide a range of cases as it is under L_1. Loss function L_1 puts extra weighting on the tails where, given a sensible choice of f, all of $f(t,\hat{\theta})$, $f(t,\hat{\theta}_t)$ and $g(t)$ will take very small values and therefore there will be little scope for improvement by using the semi-parametric method. However nearer the modes of f and g there will be more data, there may be a greater difference between the values of $f(t,\hat{\theta})$, $f(t,\hat{\theta}_t)$ and $g(t)$, and thus more scope for the semi-parametric method to improve our estimate. Since the MISE is evenly weighted it will pick out this improvement to a relatively greater extent than L_1 will.

However it must be remembered that the semi-parametric method is never going to be uniformly better than parametric estimate, which will always cross the true distribution at at least one point (see example 1, section 2.5.1).

While the condition requiring distribution f being such that $c(t,\theta)$ equals 1 may seem quite restrictive, it does at least encompass the ubiquitous Normal distribution. For other choices of f and g, I have taken a large random sample from a distribution g, then used a computer package to obtain values of $\tilde{\theta}_t$ by maximising the local likelihood function (for a wide range of target points t) and $\hat{\theta}$ for my semi-parametric and parametric estimates $f(t, \tilde{\theta}_t)$ and $f(t, \hat{\theta})$ respectively. It is then possible, by integrating numerically and using ones knowledge of what the true distribution really is, to calculate sample values for C_1, C_2 and C_3.

None of the results that I obtained required any approximations and all suggested that the semi-parametric method was superior for a far wider range of distributions and with less restrictive assumptions about h and n than in the cases we have examined so far. Indeed no counter example emerged where, when h was chosen to be large, the parametric method gave better results. They also suggested that as h decreased from a large value, the semi-parametric method increased in accuracy and thus showed even greater improvement over the parametric method. This is discussed further and some of these 'tests' are given in examples in the next section. We must remember the asymptotic nature of the results in sections 2.2 and 2.3, and that they require taking the **expectation** 'over the true distribution'. They neither prove or disprove any advantage of semi-parametric estimation in cases where h and n are not large, and nor do they prove that it will **always** give more accurate results when h and n are large.

In practical use, we would expect the semi-parametric method to perform less well than parametric estimation when for all t, $f(t,\theta) = g(t)$. Then as $n \rightarrow \infty$, the parametric estimate $f(t,\hat{\theta})$ will fit the data perfectly and cannot be bettered. The variability of any sample will cause $\hat{\theta}_t$ to deviate from $\hat{\theta}$ along the range of t and so the semi-parametric estimate will be less accurate. It thus follows that when working with finite samples, the semi-parametric estimate may also be worse if $f(t,\theta)$ and $g(t)$ are very close to one another for all t.

2.5 Case studies

In examples 1 and 3, we have generated large random samples ($n = 2000$) from distributions, chosen a large value of h ($\simeq 5$ sample standard deviations of the data) to use in our semi-parametric estimate, and considered the case where, with no knowledge of the source of the data, we incorrectly believe it to be from another parametric family. I have limited the examples to cases where this is a reasonable mistake to make - previous tests have confirmed what one would expect; that in cases where our parametric guess $f(t,\theta)$

differed more wildly from the true density function $g(t)$ with very large local differences in shape and size, the semi-parametric method had an even more substantial advantage in accuracy over the parametric method than in the more plausible cases below. However it is unlikely that initial analysis of a large sample from g will suggest that it comes from a distribution with a totally different shape and form to $g(t)$! A preparatory viewing of a histogram or kernel estimate from the data before selecting f should generally give a rough indication of the shape of the true density function $g(t)$.

Since we know $g(t)$, in these cases we can assess the performance of the parametric and semi-parametric methods of estimation, both subjectively by considering the graph and numerically by calculating the difference in performance under the three loss functions as explained in section 2.4. When using loss function L_2, the Kullback-Leibler distance, it was necessary to scale our semi-parametric estimate to ensure that it was a legitimate probability density function. It was multiplied by a constant such that it integrated to the same value as the parametric estimate over the equivalent range.

For example 2 we have no explicit density function to represent the true distribution; however it is useful as an illustration of a 'real data' case where applying the semi-parametric method gives additional accuracy in regions of t

in which we suspect that the true distribution may differ from our parametric family in shape.

When considering these examples it is important to remember that h is large and so the parametric and semi-parametric estimates will be very similar. As stated in section 2.4, examples have shown that the semi-parametric method improves substantially upon the parametric for a smaller choice of h though no theory has been produced to confirm this. An illustration of this improvement is given in example 1, plot 2e. We can also use these examples to demonstrate that the semi-parametric method does not offer uniform improvement for all t (see plot 2f).

2.5.1 Example 1

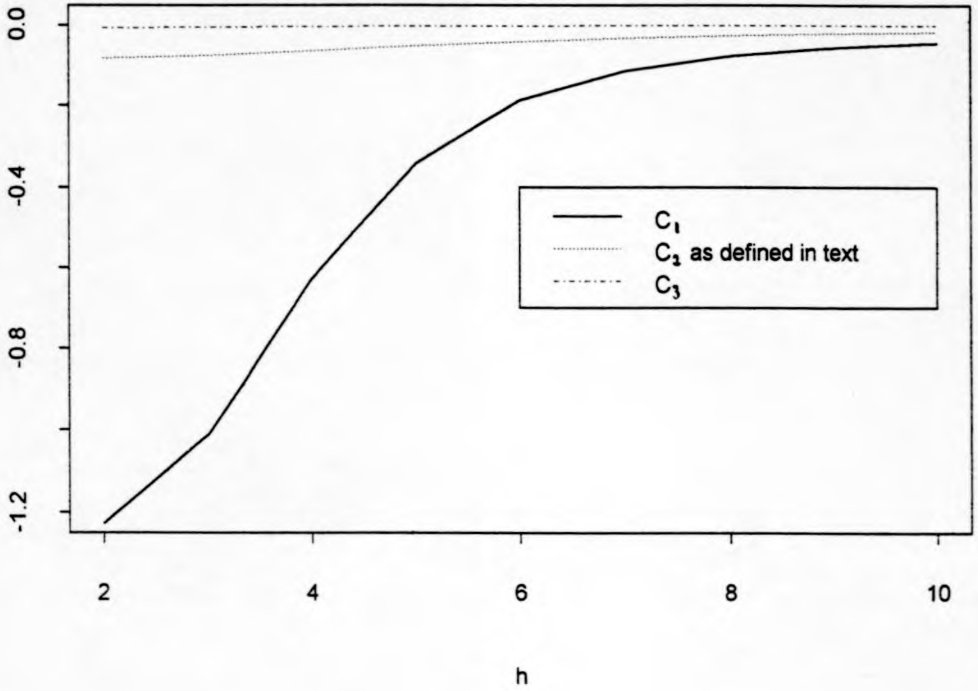
We have taken a random sample of 2000 points from a Gamma[1,4] distribution.

We now believe that the data are distributed Normally, thus our parametric estimate will be the probability density function of a Normal distribution with both mean and variance approximately equal to 4. The largest differences between this and the true distribution occur to the left of the mode and in the left tail. In these regions the local influence of the data has caused the resulting semi-parametric estimate to move away from the parametric estimate towards the true density (see plot 2d). We can use the difference C^* between the squared errors of the two methods i.e.

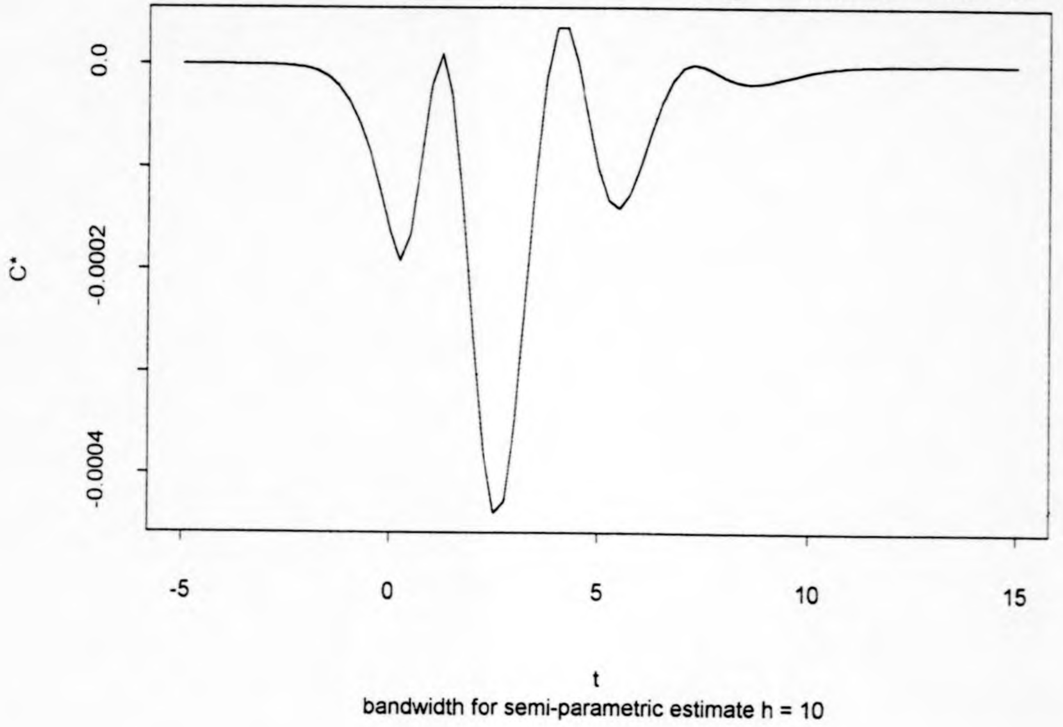
$$C^* = (f(t,\tilde{\theta}_t) - g(t))^2 - (f(t,\hat{\theta}) - g(t))^2$$

to identify the location and magnitude of our gains and losses under the semi-parametric method (see plot 2f and compare with plot 2d). This shows that the semi-parametric estimate has moved from the parametric estimate towards the true density in most areas, especially in the left tail and around the mode of $g(t)$, but this behaviour is not uniform, with the parametric method being closer to $g(t)$ (and thus pushing C^* above 0) in small regions around where it crosses the true density function. All three loss functions indicate that the semi-parametric estimate is more accurate for the majority of the range of t, the extra weighting of the L_1 loss function, particularly in the left tail, accounting for its larger value. While $L_2(f(t,\tilde{\theta}_t), g(t))$ and $L_2(f(t,\hat{\theta}), g(t))$ are not defined for $t \leq 0$, equation (13) is, taking a value of 0 in the region $t \leq 0$. Thus we can use the Kullback-Leibler distance over the whole range, though we cannot pick out numerically the advantage of the semi-parametric method in the left tail.

PLOT 2e: Comparing parametric and semi-parametric density estimation as h decreases, with g~Gamma, f~Normal



PLOT 2f : The difference between squared errors of parametric and semi-parametric estimates where g~Gamma, f~Normal



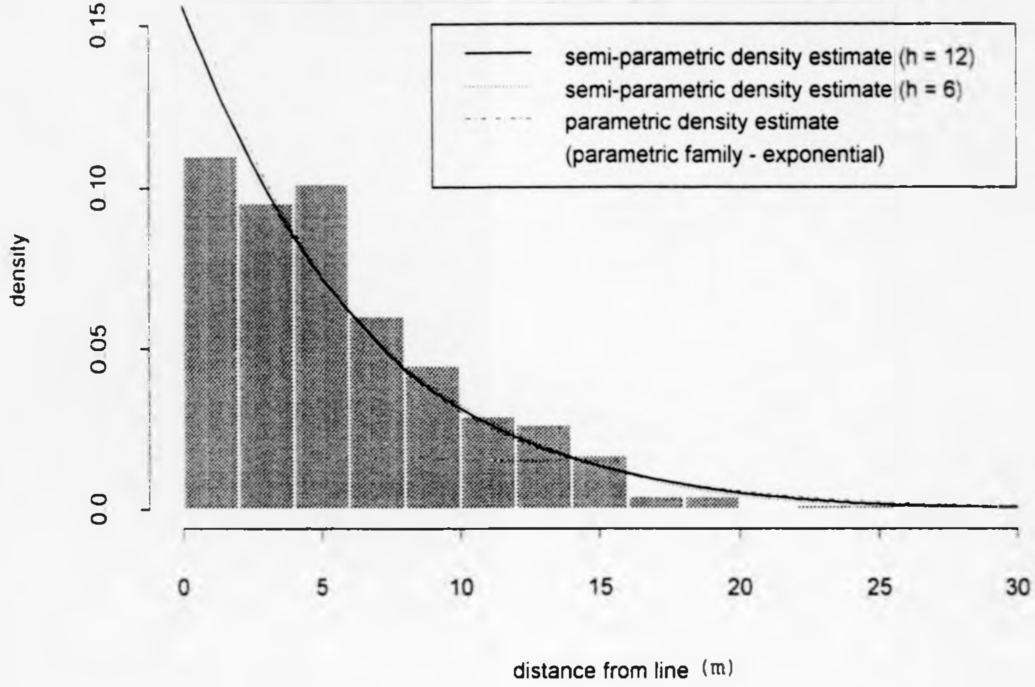
Plot 2e shows the aforementioned improvement in the accuracy of the semi-parametric method as h decreases from large values (i.e. from around 5 sample standard deviations to around 2 sample standard deviations) which allow only the very small amount of local data influence assumed in the theory of this chapter, by plotting C_1, C_2 and C_3 all against h, with loss functions calculated as described in Section 2.4.

2.5.2 Example 2

As introduced in section 1.5, the data used is a sample of 350 line transect measurements. In this case the distances are taken from the observation line to the deer that were sighted. An exponential fit appears reasonable, but we might expect the true distribution to flatten out very close to the mode (since one is as likely to see a deer one metre away as two metres away). The histogram featured on plot 2g supports this theory.

While much better results have been achieved for this data set when using a smaller h value, even using large values of h our estimate is slightly more realistic as we approach the mode. At $t = 0$, the importance of which was outlined in section 1.5, we find $f(0, \hat{\theta}_t) < f(0, \hat{\theta})$ for $h = 12$, which is around 2.5 standard deviations. Since the parametric method substantially

PLOT 2g: Normalised histogram, parametric and semi-parametric estimates of density of deer line transect data



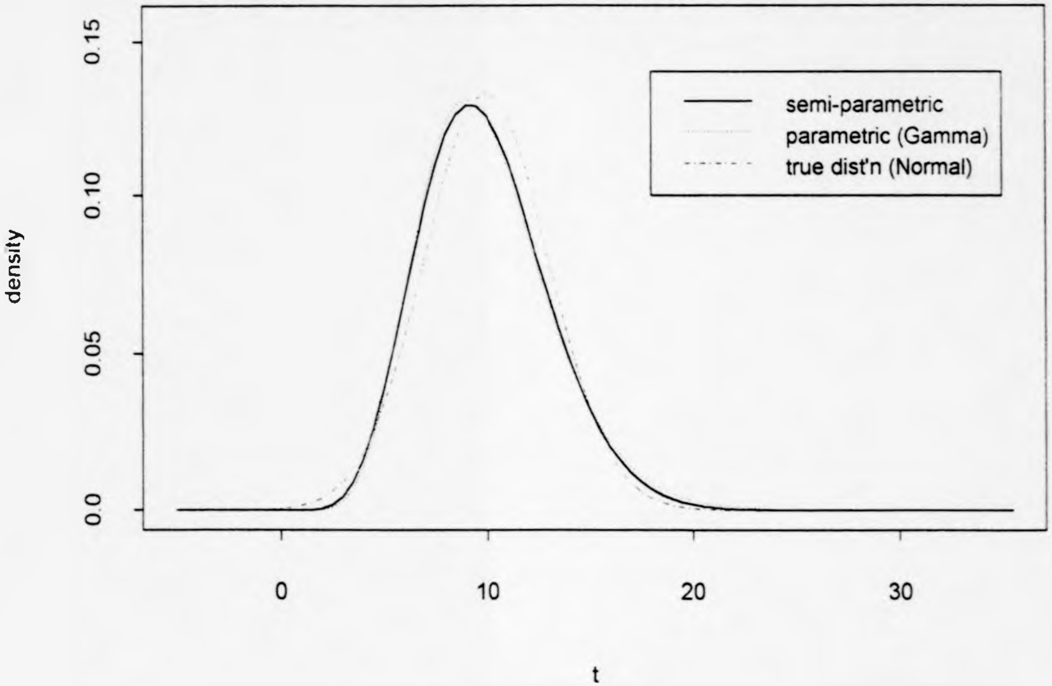
overestimates the density at this point, we have gained in accuracy by using the semi-parametric method. When a smaller value of $h = 6$ is applied, the local improvement appears to be greater (see plot 2g). This example is discussed further in forthcoming chapters.

2.5.3 Example 3

Example 3 illustrates our suggestion in section 2.4 that the semi-parametric method may well be superior to the parametric method under more general conditions than we have been able to prove in this chapter. Here f is not a two parameter family 'matching the mean and variance' as before. We have taken a sample of 2000 points from true distribution Normal[10,9] and wrongly modelled these data as from parametric family f, where $f \sim \Gamma[A, M]$. Smoothing parameter h is taken equal to approximately 3 sample standard deviations.

Our semi-parametric estimate again moves away from the parametric estimate towards the true density in several regions and is coincident with the parametric estimate elsewhere (see plot 2h). The calculations reflect this with the semi-parametric method once again performing better under all 3 loss functions considered.

PLOT 2h: Comparing parametric and semi-parametric density estimates under three different loss functions



bandwidth h=10, C_1= -0.05467391, C_2= -0.01327192 C_3= -0.0005956029

3 Introducing the 'adaption' parameter

3.1 The motivation for a variable bandwidth

Having considered large h approximations to the semi-parametric method outlined in chapter 1, we now return to exploring the more general advantages and problems of this method compared to those of parametric and non-parametric estimation in their own right, which we touched upon in section 1.5.

Ordinary kernel density estimation struggles to find a balance between achieving an accurate estimate in areas of high density and avoiding noise in the tails. Our choice of bandwidth h is central to this dilemma. We often find that in order to smooth the tails, we must accept a degree of oversmoothing and therefore underestimation of the true density elsewhere.

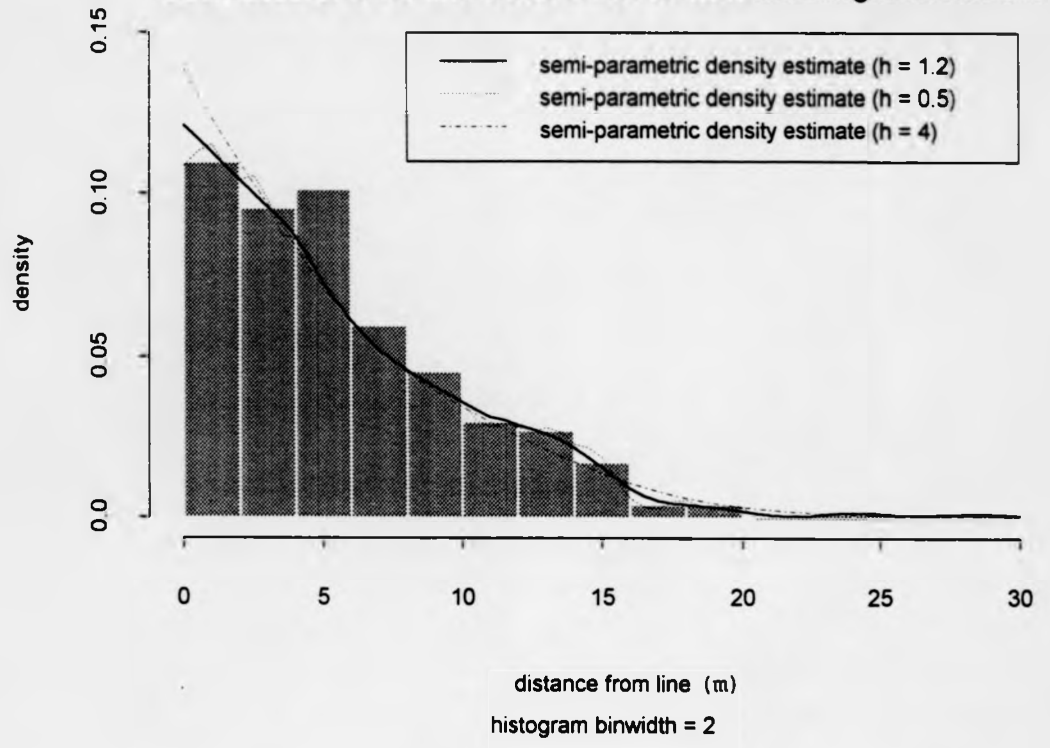
The rigidity of parametric estimation is both its principal advantage and drawback. Once we have chosen a parametric family from which we believe the data to have come, we have a simple structure to work with. It allows the incorporation of prior belief and requires only the estimation of a finite, usually small number of parameters in order to estimate the true distribution at any point t. However problems can occur in the initial step, since while many data sets appear to emanate from a particular parametric family, it is unlikely that this models the true distribution g perfectly. For example, the histogram of the deer line transect data (plot 1b) suggests that the distances from the line are distributed exponentially, except for a small region around $t = 0$ where the density flattens off. If we fit an exponential family to this data, our parametric estimate of the true density will be accurate in most regions, but will overestimate $g(t)$ as t approaches zero. Parametric estimation lacks the responsiveness to the data local to our target point t, which makes non-parametric methods very useful.

To an extent, our semi-parametric method allows us to solve the problems posed by both parametric and non-parametric estimation. We can fit a smooth parametric curve to the data and use bandwidth h, which smooths the weight function driving the local likelihood function, to determine how much influence the data has locally. As h increases this local influence will decrease. If h is chosen small and the data suggests that g differs substantially from our chosen parametric family f, then this will be reflected by the values of $\hat{\theta}_t$, which maximise the local likelihood at t, differing substantially from the MLE $\hat{\theta}$. Yet if our true distribution resembles a member of the parametric family in most areas but departs substantially from it in shape and size in small regions, then we find ourselves performing a balancing act similar to that necessary when selecting the bandwidth for ordinary kernel estimation. Consider again the line transect data. If we choose h large thus producing the smooth exponential curve which we believe to be an accurate representation of the true distribution in the tails, then there is not enough local influence of the data to pick out much of the flattening of the true density near $t = 0$ though, as illustrated in plot 2g, it is a clear improvement over the parametric estimate. Yet if we decrease h to produce the latter effect with a more non-parametric estimate, then the estimate in the right tail is no longer smooth and suffers from unwelcome local influence! Plot 3a shows this graphically, with compromise value $h = 1.2$ overestimating the mode slightly and failing to smooth the right tail completely.

The logical solution to this problem is to vary the weighting with respect to t, and therefore the amount of influence given to the data local to our target point t. This can be achieved by making bandwidth h a function of t, therefore changing the ordinary kernel $K(u)$ which we use as our weight function

$$w(x, t, h) = w(x) = K\left(\frac{x - t}{h}\right)$$

PLOT 3a: Normalised histogram and semi-parametric density estimates of deer line transect data with differing bandwidths



to a local kernel function, such that

$$w(x,t,h) = w(x) = K\left(\frac{x-t}{h_t}\right). \tag{1}$$

At any fixed t, when h_t is large we approximate a parametric estimate, and as h_t decreases to zero we now move towards the non-parametric kernel estimate $\hat{g}(t)$ of $g(t)$. We want h_t to decrease in areas where the shape of $f(t,\theta)$ differs from $g(t)$ and to make it large elsewhere. This leads to the next question; how do we formulate h_t?

3.2 Incorporating the local kernel function into semi-parametric estimation

The local kernel function is just one member of the family of adaptive or **variable** kernels, which is characterised by the non-constant nature of the bandwidth. Variable kernel functions can be divided into two distinct sub-families, depending on whether our variable bandwidth is directly related to data points $X = (x_1, ..., x_n)$ or target point t.

Ordinary kernel estimation places kernel functions with constant bandwidth h on each data point. However **varying kernel estimation** (as introduced in Chapter 1 and outlined by Silverman (1986); he defines this method as adaptive kernel estimation) allows the bandwidth to vary from one data point to the next. On the other hand, **local kernel estimation** described in Jones (1990) varies the bandwidth between target points at which we are estimating, so that given a fixed target point t, we place a kernel function with bandwidth h_t over each data point. At any one target point t, the local method gives the same estimate of $g(t)$ as the ordinary kernel method would with $h = h_t$. The local kernel density estimate over all t is actually a continuum of ordinary kernel estimates with different bandwidths for each value of t. For more detailed discussion and references on these two types of variable kernel, refer back to section 1.4.

We cannot logically insert the varying kernel into our local likelihood formula. This requires a fixed structure before the data are inserted, and if we use the varying kernel then the censoring process incorporated in the local likelihood function becomes directly dependent on the location of the data. Secondly, the motivation for using a variable kernel is to be able to vary the degree of censoring relative to the **location** at which we are estimating. It thus follows that we want the bandwidth to vary with each target point t rather than with the data points x_i, implying that our bandwidth must be a function of t. Therefore, as suggested, we use the local kernel function, defined in (1), as the weight function driving the local likelihood.

We are now required to choose a 'local' bandwidth h_t for each value of t. Jones (1990) uses the fact that at any one point t, the local method is the same as the ordinary kernel estimate with $h = h_t$. He suggests that we should proceed in a similar manner to the ordinary kernel case, by choosing h_t to minimise the MSE of the kernel estimate at t.

Splitting the MSE up into a sum of squared bias and variance, and using small h and large n approximations, we find our optimal choice of smoothing parameter is

$$h_t = \left(\frac{\int_u K(u)^2 du \; g(t)}{\left(\int_u u^2 K(u) du\right)^2 g''(t)^2 n} \right)^{-\frac{1}{5}}, \tag{2}$$

which is identical to that for the ordinary kernel estimate at t.

This method would maximise the accuracy of the local kernel estimate as an estimate of the true distribution, but we are using it as a weighting procedure and thus fine accuracy to the true distribution is of less importance. Ease of calculation is an area of concern, and (2) requires evaluation of the true density $g(t)$ and its second derivative $g''(t)$, which have to be estimated. The problems of estimating second derivatives are discussed more fully in chapter 4, but at this stage it is enough to say that it is fraught with difficulties, especially when $g(t)$ is a dramatically non-Normal, discontinuous or bounded function. For example, the exponential case poses complications around $t = 0$, where a discontinuity exists.

A preferable method of choosing h_t is suggested by the varying kernel density estimation method briefly discussed earlier. The following idea is well suited to our reasons for introducing a non-constant bandwidth into the weighting procedure. In regions of high density, where we will have lots of data helping to give an accurate kernel estimate, we want the bandwidth to take small values (and therefore cause the semi-parametric estimate to err towards non-parametric kernel estimation). There is also greater scope for a large difference between $f(t, \theta)$ and $g(t)$, and a subsequent advantage of our semi-parametric method over the parametric estimate in these areas. In regions of low density we would prefer h_t to take large values, so that the semi-parametric method approximates parametric estimation. Density functions $f(t, \hat{\theta})$ and $g(t)$ will both be small, and can therefore differ little in shape and size. There will be few data points here, making the smooth, largely parametric estimate preferable to a more non-parametric type estimate, which is liable to noise.

The varying kernel method places a kernel function over each data point

x_i, with bandwidth $h_{x_i,\alpha}$ chosen as

$$h_{x_i,\alpha} = h\left(\frac{g(x_i)}{\lambda}\right)^{-\alpha} \qquad 0 \leq \alpha \leq 1 \tag{3}$$

where

$$\lambda = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log g(x_i)\right).$$

(See Silverman (1986), chapter 5). I propose an adjustment of these formulae to make them functions of t, so that at any point t, the suggested local bandwidth is

$$h_{t,\alpha} = h\left(\frac{g(t)}{\lambda}\right)^{-\alpha}, \qquad \alpha \geq 0, \tag{4}$$

with

$$\lambda = \exp\left(\int_t g(t)\log g(t)dt\right)$$

Our local bandwidth of $h_{t,\alpha}$ is now a function of three variables; the location t and two parameters h and α. Here h can be thought of as the baseline smoothing parameter or **overall bandwidth**, controlling the underlying level of smoothing. A separate parameter α controls the extent to which $h_{t,\alpha}$ varies locally from h. **For the remainder of this thesis, we define the amount of *adaption* applied as the size of adaptive parameter α.** The direction in which $h_{t,\alpha}$ varies from h is determined by the density at t;

if $g(t)$ is less than its geometric mean λ then $h_{t,\alpha} > h$, and similarly $g(t) > \lambda$ implies $h_{t,\alpha} < h$. When α is large we find that $h_{t,\alpha}$ will decrease dramatically as $g(t)$ increases and vice-versa. As $\alpha \rightarrow 0$, the effect of adaption becomes less. When $\alpha = 0$, we have $h_{t,\alpha} = h$ for all values of t, and $w(x)$ is simply a scaled ordinary kernel function as before.

Our choice of bandwidth given in equation (4) differs from that used in varying kernel density estimation (3), both in its dependence on t rather than on data x_i, and on the removal of the upper-bound of 1 on α. There appears no reason for any specific upper-bound on α, though the semi-parametric method performs poorly if α is too large, as it would for very small h. If α is raised too high we find that the increasingly large $h_{t,\alpha}$ values in the tails makes our estimate approximate the parametric estimate very closely, whereas in regions of high density, the resulting non-parametric estimate produced by the very small $h_{t,\alpha}$ becomes just a series of spikes at the data points.

However it is logical to retain the lower bound of 0 on α given in (3). Negative values of α will lead to non-parametric tails and possibly a poor parametric estimate around the mode. The former is more serious, since if $f(t, \hat{\theta})$ is defined on an unbounded range of t, then as it becomes small the monotonically decreasing $h_{t,\alpha}$ will produce spikes at the data points and a density estimate $\simeq 0$ elsewhere. Hence α should always remain greater than or equal to 0.

3.2.1 Determining a pilot estimate

Our suggested choice of $h_{t,\alpha}$ requires the calculation of $g(t)$. Since true distribution g is of course unknown, we require a pilot estimate $\bar{g}(t)$ of the true density at t. We are not seeking fine accuracy from this estimate (if that was always possible then there would be no need for the semi-parametric method!); rather an indication of whether t is in an area of high or low density. When $\bar{g}(t)$ is used in the varying kernel, Silverman recommends use of the nearest neighbour or ordinary kernel methods for this purpose when applying the varying kernel method, but such non-parametric methods have significant drawbacks when used in a local kernel situation. These stem both from the difficulties in finding λ, the geometric mean of the pilot estimate, and from the need to find $g(t)$ over a very large range of points rather than just at the data points. For the varying kernel we calculate the sample geometric mean over the data points so that it is proportional to a sum of logarithms of the n non-parametric estimates. The local kernel method differs in that function $h_{t,\alpha}$, unlike $h_{x_i,\alpha}$, is continuous and we calculate $\bar{g}(t)$ for all t. So instead of a discrete approximation to the geometric mean, we would have to integrate over t, giving

$$\lambda = \exp\left(\int_t \bar{g}(t)\log\bar{g}(t)dt\right). \tag{5}$$

As long as it is a reasonable representation of the shape and size of $g(t)$, our priorities when choosing a pilot estimator are ease and rapidity of calculation. Thus the difficulty in calculating (5) when $\bar{g}(t)$ is found by non-parametric estimation of some sort is unwelcome. At best this value can be evaluated by numerical integration, but it cannot be found analytically. Using a kernel estimate for $\bar{g}(t)$ also creates another problem, namely how to choose a bandwidth for this estimate. Poor bandwidth selection in this situation could cause a breakdown in the whole semi-parametric process. For example, if we choose too small a bandwidth, then $\bar{g}(t)$ will become a series of spikes at the data points and will equal zero elsewhere. This in turn will produce small $h_{t,\alpha}$ values at the data points, and very large ones when $t \neq x_i$. Our eventual semi-parametric estimate will be identical to a parametric estimate except for a series of spikes at the data points. This is an interesting representation of data especially for a small sample, but a poor density estimate!

This method of finding a pilot estimate is unsatisfactory in the local kernel context. Instead I suggest taking

$$\tilde{g}(t) = f(t, \hat{\theta}),$$

the parametric estimate of the true density function $g(t)$. Presuming that our choice of parametric family f was made after consulting a histogram of the data, or at least indicates the approximate size and shape of $g(t)$, this is an effective pilot estimate. It is simple to calculate for all t, and

$$\lambda = \exp\left(\int_t f(t, \hat{\theta}) \log f(t, \hat{\theta}) dt\right) \tag{6}$$

can be calculated analytically for many choices of parametric family f.

This method has proved successful in all situations. For example, in the line transect case, even though fitting an exponential distribution to the data fails to pick out the flattening at the top (near $t = 0$) of the true density function $g(t)$, it will produce a large pilot estimate in this region, reflecting that the true distribution is similar in shape to an exponential distribution. If adaption is applied, small $h_{t,\alpha}$ values will follow here, and therefore the semi-parametric method will move towards a more non-parametric estimate around $t = 0$. This will pick out the flattening of the true density around 0. Our final estimate will be accurate to the data here and have a smooth parametric estimate in the tails where $f(t,\hat{\theta})$ is small causing $h_{t,\alpha}$ to be large. Plot 3b shows this improvement, with handpicked values of $h = 3$ and $\alpha = 1.2$. The selection of parameters h and α is discussed in the next section and some ideas for optimal selection are suggested in chapter 4.

From this point onwards we define our **adaptive semi-parametric estimate** of the density at target point t as $f(t,\hat{\theta}_{t,\alpha})$, where $\hat{\theta}_{t,\alpha}$ is the MLLE of θ calculated by maximising the local likelihood function at t, with weighting controlled by smoothing parameter $h_{t,\alpha}$. When $\alpha = 0$, this is equivalent to the ordinary semi-parametric estimate $f(t,\hat{\theta}_t)$.

3.3 Limiting properties of $h_{t,\alpha}$

We now examine the limiting behaviour of the function $h_{t,\alpha}$ with respect to its three arguments h, t and α. Without loss of generality we shall assume our data set is $X = (x_1, ..., x_n)$ with n finite. The simplest limiting properties are those corresponding to our baseline h value. If t and α are fixed, $h_{t,\alpha}$ follows the behaviour of h as it decreases to zero or increases to infinity. For all t and α,

$$h \rightarrow \infty \Rightarrow h_{t,\alpha} \rightarrow \infty$$

and

$$h \rightarrow 0 \Rightarrow h_{t,\alpha} \rightarrow 0.$$

Now taking t and h to be fixed, we consider the limits of $h_{t,\alpha}$ with respect to α. When $\alpha = 0$ we have the ordinary semi-parametric method with $h_{t,\alpha} = h$ everywhere. As α increases from 0 the behaviour of $h_{t,\alpha}$, and thus our estimate, depends on the location of t. Define

$$\Upsilon(t) = \frac{f(t, \hat{\theta})}{\lambda}, \tag{7}$$

which is constant with respect to α. Then if $\Upsilon(t)$ is greater than 1,

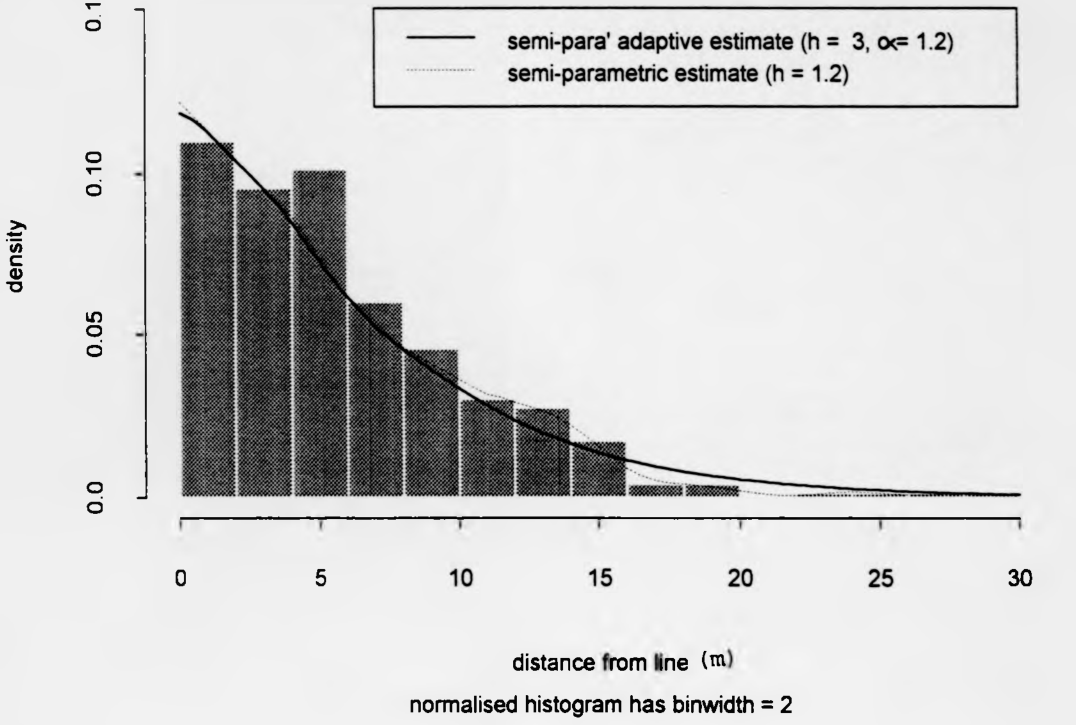
$$h_{t,\alpha} = h \exp(-\alpha \log \Upsilon(t))$$

will decrease towards zero as α increases, the rate dependent on the value of $\Upsilon(t)$. Alternatively if $\Upsilon(t)$ is less than 1, an increase in α will send $h_{t,\alpha}$ towards infinity. If t is such that $\Upsilon(t) = 1$, then applying adaption will have no effect on our estimate, the amount of smoothing being totally governed by the size of h. We define the set

$$B^{*} = (t : \Upsilon(t) = 1) \tag{8}$$

as the set of **Boundary Points**, since they are the values on the boundary between $h_{t,\alpha}$ becoming greater or less than h as α increases. Using the Mean

PLOT 3b: Comparing adaptive and non-adaptive semi-parametric density estimates of deer line transect data



Value Theorem, it is trivial to show that B^* is non-empty for any probability density function $f(t,\theta)$.

The limiting behaviour with respect to t is itself dependent on the function $f(t,\hat{\theta})$. Given fixed h and α, then since $\Upsilon(t)$ is bounded above by

$$\max_t \Upsilon(t) = \frac{\max_t f(t,\hat{\theta})}{\lambda},$$

$h_{t,\alpha}$ will have a greatest lower bound which is positive and attained at the mode. The existence of an upper bound depends on the range of t on which $f(t,\theta)$ is defined. If this is bounded at both ends (for example, when f is a uniform distribution), and a minimum value ¿ 0 is attained by $f(t,\theta)$, then $h_{t,\alpha}$ will have a least upper bound when h and α are fixed. However, if the range is unbounded at either end, we can always find t such that $f(t,\hat{\theta})$, and therefore $\Upsilon(t)$, can take a value arbitrarily close to zero. Then, as $\Upsilon(t) \rightarrow 0$, we get $h_{t,\alpha} \rightarrow \infty$.

3.4 Quantifying the improvement offered by adaption

In the example above we can visually observe the improvement caused by applying adaption. But it is not clear whether applying adaption always improves our estimate, and when it does offer an improvement, how much of one is possible. We can examine these questions in two ways. Initially we shall consider extending the large h approximations of chapter 2 to large $h_{t,\alpha}$ approximations, and observing the effect of increasing the amount of adaption α very slightly from 0. Alternatively we can assume $\alpha > 0$, and ignore the tail regions where $h_{t,\alpha}$ will become very large as $\alpha \rightarrow \infty$, causing $f(t, \hat{\theta}_{t,\alpha})$ to approximate $f(t, \hat{\theta})$. Instead we concentrate on the target points t for which $h_{t,\alpha} < h$ for all $\alpha > 0$, where adaption could dramatically change our estimate from being largely parametric to being largely non-parametric.

3.4.1 Some approximations when $h_{t,\alpha}$ is large

We assume that $h_{t,\alpha}$ large for all t, which in turn places a restriction on our examination of the effects of adaption. As α increases towards infinity, $h_{t,\alpha}$ will decrease from h towards zero in areas where the pilot estimate is large, contradicting any large $h_{t,\alpha}$ assumptions. So as well as assuming our baseline h value to be large, we also restrict our considerations on the effect of adaption to the effect of a small increase of α from zero, such that $h_{t,\alpha}$ is still large everywhere. Given a fixed overall bandwidth h, we cannot consider evaluating the performance of $f(t, \hat{\theta}_{t,\alpha})$ as α increases further.

Taking

$$\delta_{t,\alpha} = \frac{1}{h_{t,\alpha}},$$

the small δ approximations of Copas (1995a) and chapter 2 can simply be changed to small $\delta_{t,\alpha}$ approximations. Replacing $\delta = \frac{1}{h}$ by $\delta_{t,\alpha}$, our weight function can be approximated by

$$K\left(\delta_{t,\alpha}(x_i - t)\right) = 1 + \frac{1}{2} b_1 \delta_{t,\alpha}^2 (x_i - t)^2 + O(\delta_{t,\alpha}^4), \tag{9}$$

where $b_1 = K''(0)$. Our adaptive local likelihood score can be written such that

$$\frac{d}{d\theta} L_w(X, t, \theta) \simeq \frac{d}{d\theta} L(X, t, \theta) + \frac{1}{2} b_1 \delta_{t,\alpha} n T. \tag{10}$$

Then, for example, we have

$$(\tilde{\theta}_{t,\alpha} - \hat{\theta}) \simeq -\frac{1}{2}\delta_{t,\alpha}^2 \left(E_f \left(\frac{d^2}{d\theta^2} \log f(X, \theta) \right) \right)^{-1} T$$

and this leads to

$$E_g(\tilde{\theta}_{t,\alpha} - \hat{\theta}) \simeq -\frac{1}{2}\delta_{t,\alpha}^2 I(\theta)^{-1} \int_x (t - x)^2 \frac{d}{d\theta} \log f(x, \theta)(g(x) - c(t, \theta) f(x, \theta)) dx, \tag{11}$$

where

$$c(t, \theta) = \frac{\int_x (t - x)^2 g(x) dx}{\int_x (t - x)^2 f(x, \theta) dx},$$

$$I(\theta)^{-1} = b_1 \left(E_f\left(\frac{d^2}{d\theta^2}\log f(t,\theta)\right)\right)^{-1}$$

and

$$T = n^{-1}\sum_{i=1}^{n}(x_i - t)^2\left(\frac{d}{d\theta}\log f(x_i,\theta) - \frac{d}{d\theta}\log(\sigma_\theta^2 + (t-\mu_\theta)^2)\right)$$

as before.

3.4.2 Examining the effect of a small amount of adaption

We can now determine whether increasing α from 0 will automatically improve our estimate. As in chapter 2 we will use the Kullback-Leibler distance to measure the accuracy of our estimates to the true distribution. Consider the risk function C_2 introduced in subsection 2.3.2. This was the expected difference between the Kullback-Leibler distance of our semi-parametric estimate $f(t,\hat{\theta}_t)$ from $g(t)$ and the Kullback-Leibler distance of the ordinary parametric estimate $f(t,\hat{\theta})$ from $g(t)$. A negative value of $E(C_2)$ indicated that we would expect the semi-parametric estimate to be more accurate.

Define $C_{2,\alpha}$, the adaptive analogue to C_2, as

$$C_{2,\alpha} = L_2(\bar{f}(t,\hat{\theta}_{t,\alpha}), g(t)) - L_2(f(t,\hat{\theta}), g(t)),$$

where $\bar{f}(t,\hat{\theta}_{t,\alpha})$ is the normalised version of $f(t,\hat{\theta}_{t,\alpha})$ and loss function L_2 is the Kullback-Leibler distance between two density functions, first defined in chapter 2, equation (18).

To progress we will need to use the large $h_{t,\alpha}$ approximations given in the preceding subsection, which are really small $|x_i - t|h_{t,\alpha}^{-1}$ approximations. When t is fixed we can always select an overall bandwidth h to fulfil the latter criterion but, as in the non-adaptive case of chapter 2, when we need to integrate over an interval of t and require small $|x_i - t|h_{t,\alpha}^{-1}$ approximations throughout the interval, such a selection of h may be impossible.

If the domain of $f(t,\theta)$ is bounded at both ends, then there is no problem because $|t - x_i|$ is then bounded. We can take h as any fixed value larger than $\max_{t,x_i} |(t - x_i)\Upsilon(t)^{\alpha^\circ}|v$, where v is a large constant and α° is the value of α up to which we shall consider our large $h_{t,\alpha}$ approximation valid. The greater the amount of α we want to be able to consider being applied, the larger we will have to choose v.

Unfortunately $f(t,\theta)$ is often defined on an unbounded interval of t, and when examining the performance of $f(t,\hat{\theta}_{t,\alpha})$ over all t we will face similar problems to those encountered in the non-adaptive case, namely that for any fixed h and data set $x_1, ..., x_n$ we can always choose t such that $|x_i - t|h_{t,\alpha}^{-1}$ is large. However because of the local nature of $h_{t,\alpha}$, which increases as $f(t,\hat{\theta})$

decreases, it maybe that these problems are limited to only certain choices of f. For example, if f is a Normal distribution, then $|t| \to \infty$ such that $|t - x_i| \to \infty$ for all x_i, but

$$|(t - x_i)f(t, \hat{\theta})^{\alpha}| = \left|(t - x_i)\left(\frac{1}{\sqrt{2\pi}\hat{\sigma}}\right)^{\alpha} \exp\left(-\frac{\alpha}{2}\left(\frac{t - \hat{\mu}}{\hat{\sigma}}\right)^{2}\right)\right|$$

which will converge to 0 as $|t| \to \infty$. Thus $|(t-x_i)\Upsilon(t)^{\alpha}|$ will be maximised at a target point within the interval $(-\infty, \infty)$ rather than diverging to infinity as $|t|$ becomes large. An appropriate choice of

$$h = v \max_{t,x_i} \left|(t - x_i)\Upsilon(t)^{\alpha^{\circ}}\right| \tag{12}$$

where v is a large constant will ensure that $|t - x_i|h_{t,\alpha}{}^{-1}$ is small everywhere for values of α up to α°.

However this does not work for all parametric pilot estimates $f(t, \hat{\theta})$. A counter example is provided by the Cauchy(0) Distribution. If

$$f(t, \hat{\theta}) = \frac{1}{\pi(1 + t^2)},$$

then

$$(t - x_i)\Upsilon(t)^{\alpha} \to \pi^{-1}\lambda^{-\alpha}t^{1-2\alpha}$$

as $|t|$ increases, which in turn diverges to infinity if α takes values less than

$\frac{1}{2}$. However large we choose h, we will always be able to find a value of t large enough such that $|t - x_i|{h_{t,\alpha}}^{-1}$ is no longer small.

This hurdle is overcome by mirroring the argument given in subsection 2.3.1 which, as explained in subsection 2.3.2, is equally valid when we are using the Kullback Leibler distance as our loss function. After replacing $f(t, \tilde{\theta}_t)$ by $f(t, \tilde{\theta}_{t,\alpha})$, this limiting argument transfers automatically to the adaptive case.

When α is equal to zero, then $C_{2,\alpha}$ is equivalent to C_2; calculating the expectation of this over the data we can refer to the results of subsection 2.3.2 which proved that you would expect the semi-parametric method to be superior under certain conditions. Because of the continuous nature of $h_{t,\alpha}$ over the range of t, it follows that, under the same conditions, we can always choose an α value greater than 0 such that we expect the adaptive semi-parametric estimate $f(t, \tilde{\theta}_{t,\alpha})$ to still be closer to the true distribution than parametric estimate $f(t, \tilde{\theta})$ under both distance measures considered in subsections 2.3.1 and 2.3.2. Therefore as $\alpha \to 0$,

$$f(t, \tilde{\theta}_{t,\alpha}) \to f(t, \tilde{\theta}_t) \Rightarrow E_g(C_{2,\alpha}) \to E_g(C_2) \leq 0.$$

However we are less interested in this than in whether applying adaption will actually improve upon our ordinary semi-parametric estimate. So instead we consider

$$\frac{d}{d\alpha}\,|_{\alpha=0}\; E(C_{2,\alpha}).$$

The Kullback-Leibler distance of the parametric estimate $f(t,\hat{\theta})$ from $g(t)$ is independent of α so this is effectively just measuring the direction of change in the distance from $f(t,\tilde{\theta}_{t,\alpha})$ to $g(t)$ as α becomes positive. The sign of this value will indicate whether increasing α slightly from zero, thus applying a small amount of adaption, will make our semi-parametric estimate more or less accurate when the overall bandwidth $h = \frac{1}{\delta}$ is large.

From the extensions from chapter 2 given in subsection 3.4.1, we can substitute $\delta_{t,\alpha}$ for δ in subsection 2.3.2, equation (19), giving

$$\frac{d}{d\alpha}E(C_{2,\alpha}) \simeq$$

$$\frac{1}{2}\int_t\int_x \left(\frac{d}{d\alpha}\delta_t\right)(x-t)^2\rho(t)^T I(\theta)^{-1}\rho(x)(g(x)-f(x,\theta)c(t,\theta))\eta(t)dxdt \quad (13)$$

where

$$\rho(x) = \frac{d}{d\theta}\log f(x,\theta),$$

$$\eta(x) = g(x) - f(x,\theta)$$

and sample size n is assumed large. The square of our inverted local bandwidth, $\delta^2_{t,\alpha}$, can be written as

$$\delta^2_{t,\alpha} = \delta^2 \Upsilon(t)^{-2\alpha},$$

and after differentiating with respect to α we can construct the following formula in terms of δ;

$$\frac{d}{d\alpha}\delta^2_t = \delta^2 \frac{d}{d\alpha}\exp(-2\alpha \log \Upsilon(t)) = 2\delta^2 \Upsilon(t)^{-2\alpha} \log \Upsilon(t).$$

Evaluating this at $\alpha = 0$ and inserting into equation (13) we get

$$\left.\frac{d}{d\alpha}\right|_{\alpha=0} E(C_{2,\alpha}) =$$

$$\delta^2 \int_t \int_x (x-t)^2 \rho(t)^T I(\theta)^{-1} \rho(x)(g(x) - f(x,\theta)c(t,\theta))\eta(t) \log \Upsilon(t) dx dt. \quad (14)$$

This integral has proved intractable when attempting to calculate it for general f and g, but it can be evaluated analytically for some specific examples, and numerically for any specific f and g. First we consider a case where we can calculate $\frac{d}{d\alpha}|_{\alpha=0} E(C_{2,\alpha})$ analytically.

Take $f \sim$ Normal, in which case $c(t,\theta) = 1$. Without loss of generality, we can assume that true distribution g has mean equal to 0 and variance σ^2. We first evaluate $\log \Upsilon(t)$ and integral (14) can then be written as

$$\left.\frac{d}{d\alpha}\right|_{\alpha=0} E(C_{2,\alpha}) \simeq \frac{1}{2}\delta^2 \int_t \int_x (x-t)^2 \rho(t)^T I(\theta)^{-1} \rho(x)\eta(x)\eta(t) \left(1 - \frac{t^2}{\sigma^2}\right) dx dt.$$

Using the working and notation of subsection 2.3.2, equation (14) reduces further to

$$-\frac{\delta^2(F_3-G_3)^2}{2\sigma^4}+\frac{1}{2\sigma^2}\delta^2\int_t\int_x t^2(x-t)^2\rho(t)^T I(\theta)^{-1}\rho(x)\eta(x)\eta(t)dxdt. \quad (15)$$

We will now work on the integral part of the second half of this sum.

$$\int_t\int_x t^2(x-t)^2\rho(t)^T I(\theta)^{-1}\rho(x)\eta(x)\eta(t)dxdt =$$

$$\int_t\int_x (t-x)^2t^2 \begin{pmatrix} \frac{d}{d\mu}\log f(x,\theta)\eta(x) \\ \\ \frac{d}{d\sigma}\log f(x,\theta)\eta(x) \end{pmatrix}^T \begin{pmatrix} 1 & 0 \\ \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \frac{d}{d\mu}\log f(t,\theta)\eta(t) \\ \\ \frac{d}{d\sigma}\log f(t,\theta)\eta(t) \end{pmatrix} dxdt$$

$$\begin{aligned}
&= \int_t t^3\frac{d}{d\mu}\log f(t,\theta)\eta(t)dt\int_x -2x\frac{d}{d\mu}\log f(x,\theta)\eta(x)dx \\
&+ \int_t t^2\frac{d}{d\mu}\log f(t,\theta)\eta(t)dt\int_x x^2\frac{d}{d\mu}\log f(x,\theta)\eta(x)dx \\
&+\frac{1}{2}\int_t t^3\frac{d}{d\sigma}\log f(t,\theta)\eta(t)dt\int_x -2x\frac{d}{d\sigma}\log f(x,\theta)\eta(x)dx \\
&+\frac{1}{2}\int_t t^2\frac{d}{d\sigma}\log f(t,\theta)\eta(t)dt\int_x x^2\frac{d}{d\sigma}\log f(x,\theta)\eta(x)dx \\
&= \frac{2}{\sigma^4}(G_3-F_3)^2+\frac{1}{2\sigma^6}(G_4-F_4)^2-\frac{1}{\sigma^6}(G_3-F_3)(G_5-F_5)
\end{aligned}$$

where G_3, G_4, G_5, and F_3, F_4 and F_5 are the third, fourth and fifth moments of g and f respectively. Our knowledge of the density function of f tells us that both F_3 and F_5 will be 0, and that $F_4 = 3\sigma^4$. Thus equation (14) is now of the form

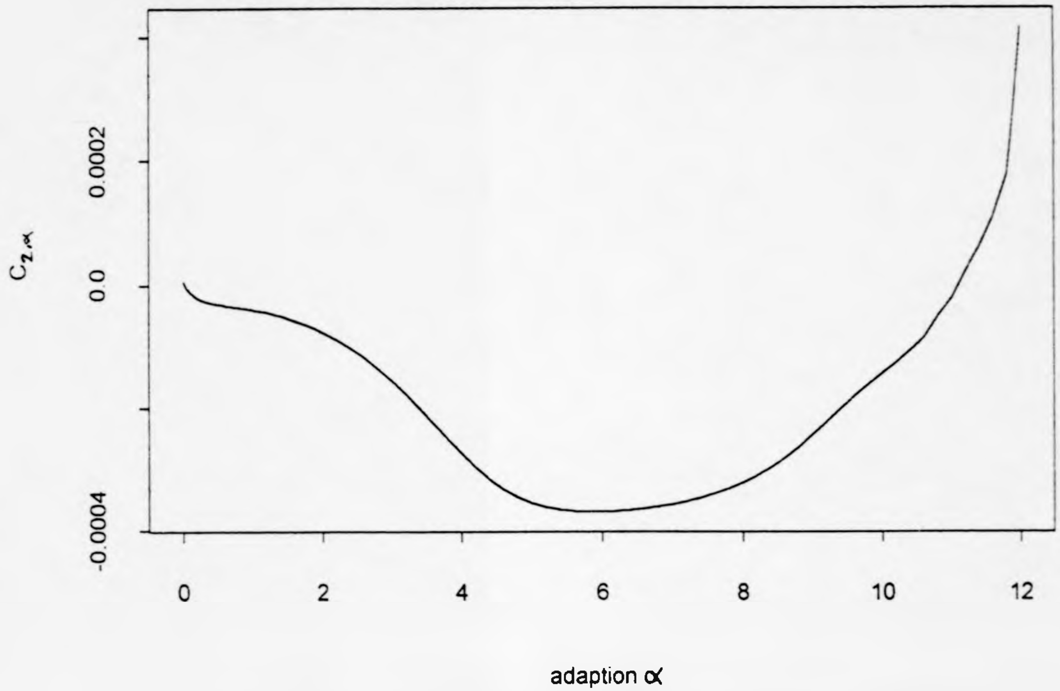
$$\left.\frac{d}{d\alpha}\right|_{\alpha=0} E(C_{2,\alpha}) \simeq \frac{\delta^2}{2}\left(-\frac{3G_3^2}{\sigma^4} - \frac{(G_4 - F_4)^2}{2\sigma^6} + \frac{G_3G_5}{\sigma^6}\right). \tag{16}$$

The sign of equation (16) determines whether increasing α from 0 will initially improve or lessen the accuracy of $f(t, \hat{\theta}_{t,\alpha})$ to $g(t)$. This is dependent on the moments of g. For example, if g is symmetric around zero, then if G_3 and G_5 exist they will be equal to 0, and

$$\left.\frac{d}{d\alpha}\right|_{\alpha=0} E(C_{2,\alpha}) \simeq -\frac{(G_4 - F_4)^2}{2\sigma^6} \leq 0.$$

Then an initial application of adaption will cause a decrease in the value of $E(C_{2,\alpha})$, indicating that $L_2(f(t, \hat{\theta}_{t,\alpha}), g(t))$ is decreasing and that our semi-parametric estimate is thus increasing in accuracy. Plot 3c illustrates this result in the form of a practical simulation. A random sample of 5000 points was taken from a bimodal distribution, which was an equal mixture of two Normal distributions with means of plus and minus 1, and both with variance 2. An adaptive semi-parametric estimate was constructed, where $f \sim$ Normal and using large $h \simeq 8$ sample standard deviations. Estimate $f(t, \hat{\theta}_{t,\alpha})$ was scaled such that it integrated to the same value as the parametric estimate over the same range. This was necessitated by our use of the Kullback-Leibler distance measure, which requires both functions to be probability

PLOT 3c: Adaption vs the gain in accuracy of adaptive semi-parametric estimation over parametric estimation FOR LARGE h



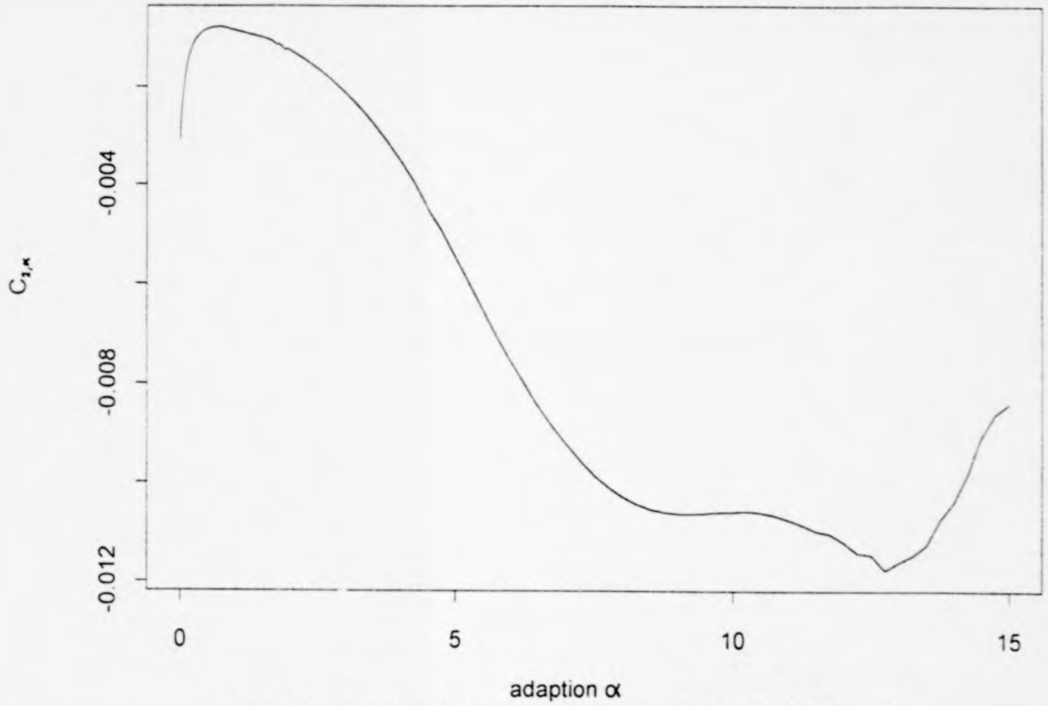
true dist'n g is bimodal mixture of Normals, chosen parametric family f is Normal density functions. Then $C_{2,\alpha}$ was estimated by numerical integration. Plot 3c shows the expected initial decrease of $C_{2,\alpha}$, reaching a minimum at $\alpha \simeq 6$ and then increasing as the very large values of α cause the density estimate around the mode to become noisy and inaccurate.

However when $g(t)$ is non-symmetric, increasing α from 0 does not guarantee immediate improvement. For example, if g is a Gamma distribution then we find that the differential of $E(C_{2,\alpha})$ with respect to α, given in (16), is greater than 0 when $\alpha = 0$. So the improvement in accuracy emanating from applying a small amount of adaption is not uniform for all $g(t)$.

This can be illustrated by calculating $C_{2,\alpha}$ by numerical integration for some large random samples from non-symmetric g, where f is assumed Normal, and then plotting against α. Plot 3d shows this result, using a random sample of 5000 points from a Gamma[1,4] distribution and fitting a Normal distribution to these points using the adaptive semi-parametric method, with the final density estimate appropriately scaled as in the previous example. Again h is large, being taken $\simeq 8$ sample standard deviations.

So it appears that the result of increasing α slightly from 0 when h and n are large is dependent on the contrasting shapes of $f(t,\theta)$ and $g(t)$. Since the extreme tails, where $\Upsilon(t) \simeq 0$, are where a very small amount of adaption

PLOT 3d: Adaption vs gain in accuracy of adaptive semi-parametric estimation over parametric estimation FOR LARGE h



true dist'n g is Gamma(1,4), chosen parametric family f is Normal will have the greatest effect, it would be logical to assume that it is the behaviour in these regions which determines the response of $C_{2,\alpha}$ to the initial application of α.

However since $\Upsilon(t)$ is so small, and $f(t,\hat{\theta}_{t,\alpha})$, $f(t,\hat{\theta})$ and $g(t)$ are all likely to be close to one another and to 0, **the effect of α in these regions compared to its effect in regions of high density** (where our parametric and true distributions may differ more substantially) will decrease rapidly as α increases. Thus in plot 3d, after α is greater than 0.3, the relationship between $C_{2,\alpha}$ and α will be dominated by regions where $g(t)$ and $f(t,\theta)$ differ more more in value; we examine this phenomenon in the next section.

3.4.3 Considering $h_{t,\alpha}$ on a restricted range

While we cannot extend the above proof to predict the behaviour of $f(t,\hat{\theta}_{t,\alpha})$ for larger values of α or for different choices of f, by restricting our examination to certain subsets of the range of t we can at least come to some useful conclusions.

In both examples above, the overall bandwidth h is chosen large enough such that our large $h_{t,\alpha}$ approximations would be valid for all t up to around $\alpha = 0.8$. However, since in these examples our adaptive semi-parametric estimate is calculated directly without reliance on any approximations, then as well as supporting our theoretical results over the large $h_{t,\alpha}$ range of ($0 \leq \alpha \leq 0.8$), we can also examine the behaviour of our $f(t, \hat{\theta}_{t,\alpha})$ outside this range. In plots 3c and 3d we see that the maximum improvement over our parametric estimate (and therefore the minimum Kullback-Leibler distance between $g(t)$ and $f(t, \hat{\theta}_{t,\alpha})$) is attained at a much larger value of α, at which our large $h_{t,\alpha}$ approximations would be invalid.

Whilst the behaviour of the two cases differs for very small α, as it increases both follow the pattern we would expect given the motivation behind introducing adaption. Having seen the minimum of $C_{2,\alpha}$ attained, the effect of a further increase in α is as foretold in section 3.2, page 76, producing increasingly noisy estimates in areas of high density. Our motivation behind the introduction of adaption and the method used was to have a more parametric estimate in regions where both distributions f and g have low density, and a more non-parametric estimate in regions where both were of higher density with greater scope for $f(t, \theta)$ to differ from $g(t)$ in shape and size. We rely on our chosen parametric family, being realistic so that the high and low density areas of f and g coincide. The choice of f is crucial since our pilot estimate $f(t, \hat{\theta})$ determines whether adaption increases or decreases

$h_{t,\alpha}$ from h.

Define the region of high density as D^*, where

$$D^* = (t : \Upsilon(t) > 1). \tag{17}$$

Thus D^* and D^{*C} partition the range of t on which $f(t,\theta)$ is defined, with the cuts of the partition being the boundary points defined by equation (8) in section 3.3. Consider D^{*C} where, as all of $f(t,\hat{\theta}_{t,\alpha})$, $f(t,\hat{\theta})$ and $g(t)$ are relatively small, increasing α and thus $h_{t,\alpha}$ will have little effect on increasing or decreasing the accuracy of $f(t,\hat{\theta}_{t,\alpha})$ to $g(t)$, especially if h is already large ensuring that $f(t,\hat{\theta}_{t,\alpha}) \simeq f(t,\hat{\theta})$. The principal action will occur in D^* where we will move towards a more locally influenced non-parametric estimate. Assume that overall bandwidth h is at least large enough to ensure that the tails of $f(t,\hat{\theta}_{t,\alpha})$ are 'parametric enough' to be smooth. Ignore the small gain or loss in accuracy that will occur in D^{*C} and concentrate solely on D^*.

If $g(t)$ differs noticeably from $f(t,\hat{\theta})$ in shape over D^*, then for nearly all $t \in D^*$ (obviously not at all t since the parametric estimate may cross the true distribution at some $t \in D^*$) a more non-parametric estimate is likely to improve our accuracy to $g(t)$. Increasing α to a value greater than 0 will be necessary to achieve this. Quite how much is needed will depend on the difference from $f(t,\hat{\theta})$ to $g(t)$ and the size of h. Consider the examples above; the shapes of the density functions defining f and g differ less in the first example, so our adaptive semi-parametric estimate has to move less towards a local kernel estimate to minimise $C_{2,\alpha}$ than it does in the second example, resulting in the smallest value of $C_{2,\alpha}$ in plot 3c occurring at a smaller value of α than in plot 3d.

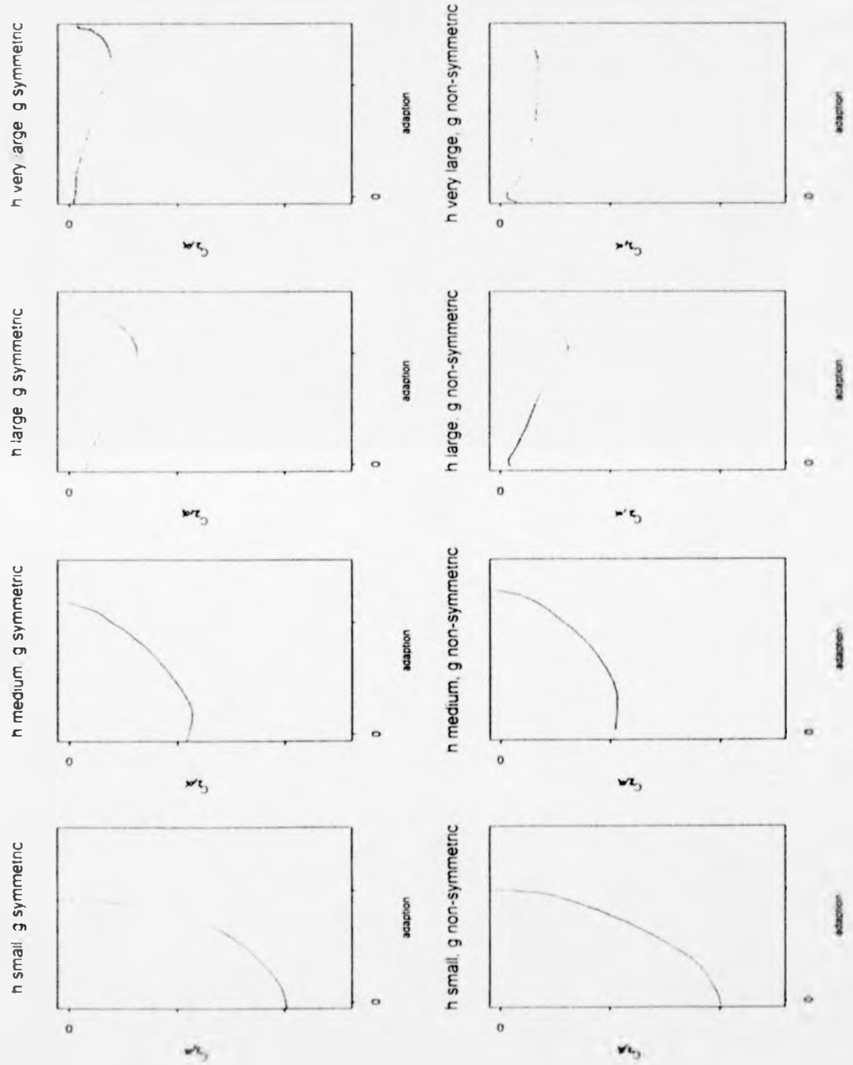
Thus when h is large there will exist a value of α in the interval $(0,\infty)$ which will maximise our accuracy in D^* and will be close to the overall optimum value of α when considering the full range of t. Devising an automatic method for choosing an optimum α using this restriction on the range of t is discussed in chapter 5, and the relationship between results over D^* and the full range of t are examined further with a few practical examples.

3.4.4 The effect of α when h is small

Having considered the behaviour of our estimate for large h as α is increased from 0 through to large α, we now turn our attention to the case where overall bandwidth h is smaller. We initially restrict ourselves to considering the regions of high density, with $f(t,\hat{\theta})$ again differing noticeably in shape from $g(t)$. As h becomes smaller, less α will be need to reduce $h_{t,\alpha}$ from h to its optimum value for all $t \in D^{*}$. So in terms of maximising accuracy over D^{*}, as $h \rightarrow 0$, our optimum choice of α will also decrease towards 0. The sketches in Plot 3e show the way we would expect the relationship between $C_{2,\alpha}$ and α to respond as h decreases from a very large value. The cases when g is symmetric and non-symmetric are both considered. However, if h is small, the effect of adaption in D^{*C} becomes significant. If $f(t, \theta)$ differs from $g(t)$ in D^{*C}, then increasing adaption may lead to a small but noticeable loss of accuracy here causing our optimum α over the whole range to be slightly smaller than the optimum choice when just considering D^{*}.

If h is so small that $f(t, \hat{\theta}_{t,\alpha})$ is a noisy non-parametric estimate with spikes at the data points, then a significant improvement in our estimate of $g(t)$ in D^{*C} will occur as α increases producing a smoother estimate. However this increase will be making our estimate in D^{*} even more non-parametric, though the upper-bound on $\Upsilon(t)$ and the greater amount of data in this region will temper the problem. Whether our optimum value of α is at 0 or in the range $(0, \infty)$ in this situation will once again depend on the difference between $f(t, \theta)$ and $g(t)$, but it is clear that having a very small overall bandwidth h is undesirable. If h is small enough to make our estimate in D^{*} too noisy before adaption is even applied, then a larger value of h should

PLOT 3e: Behaviour of $C_{2,\infty}$ vs ∝ as h varies



be chosen. In all of the above discussion we are assuming n fixed and large. Given h fixed and large, we would expect less α to be required for optimal accuracy as n decreases, since with a smaller sample size we will want larger local bandwidths $h_{t,\alpha}$ for all $t \in D^*$.

3.5 Comment

Adaption offers the possibility of significantly improving the accuracy of our semi-parametric estimate. It is at its most effective when $f(t, \hat{\theta}) \simeq g(t)$ in the tails, but differs significantly in D^*. We must now consider when to apply it and how much to apply.

4 An automatic method for choosing overall bandwidth h

4.1 Introduction

Having discussed the theory behind our semi-parametric density estimation method and the role of adaption in improving it, we now turn our attention to the more practical topic of choosing 'best' values of h and α. Section 4.2 proposes an automatic method of selecting the overall bandwidth h, while the incorporation of prior belief into this procedure is introduced in chapter 6. The method of choosing h developed in this chapter has similarities with those suggested in Silverman (1986) for choosing the bandwidth used in ordinary kernel density estimation, in that we select it to minimise the mean squared error (MSE) of our density estimate.

4.2 Choice of h

The introduction of adaption makes our choice of a suitable overall bandwidth a much easier task. It is no longer a balancing act between attaining accuracy around the mode and smoothness in the tails of our density estimate; now both of these ideals can be simultaneously achieved. Using the notation of chapter 3, $h_{t,\alpha}$ is defined for all t as

$$h_{t,\alpha} = h\Upsilon(t)^{-\alpha}, \tag{1}$$

where

$$\Upsilon(t) = \frac{f(t,\hat{\theta})}{\lambda}$$

and

$$\lambda = \exp(E_f(\log f(t,\hat{\theta}))).$$

Applying adaption will cause $h_{t,\alpha}$ to differ from h in areas where the density function of f is larger or smaller than λ. This will produce a smooth parametric density estimate in the tails and a more non-parametric kernel type estimate around the mode. Therefore when choosing an overall bandwidth for use in the adaptive semi-parametric method, it makes sense to consider the behaviour of our $f(t,\hat{\theta}_{t,\alpha})$ in the regions where $h_{t,\alpha} \simeq h$ for all α. These areas are centred around the boundary points, defined in chapter 3, equation (8), where adaption has no effect. If we choose h such that we obtain a sensible estimate in these regions, then a small amount of adaption should bring the required smoothness to the tails of our estimate. At the same time, given a reasonable choice of parametric family f, it will cause $f(t,\hat{\theta}_{t,\alpha})$ to move towards an accurate non-parametric estimate in areas where we have the greatest amounts of data.

4.2.1 Constructing the selection method

Having established a policy for choosing h, we now examine the methodology. The first requirement is to locate the boundary point mentioned above by calculating $\lambda = \exp(E_f(\log f(t, \hat{\theta})))$ and finding the value of $t = t^*$ such that $f(t^*, \hat{\theta}) = \lambda$. Parameter λ is the geometric mean of $f(t, \hat{\theta})$ over distribution f. At least one boundary point will always exist for any continuous function $f(t, \theta)$. Section 4.3 discusses how to deal with cases where more than one boundary point exists.

We then find h', where choosing $h = h'$ minimises the MSE of the semi-parametric method at the boundary point. To obtain a workable formula for the MSE, we initially assume bandwidth h is small, sample size n is large, and use the following approximations and theory from Copas (1995b). These concern the ordinary semi-parametric method without the application of any adaption. However, we are only interested in the behaviour of the adaptive semi-parametric method at the boundary points, where adaption has no effect, so this is not a problem. In the following equations, scaled kernel function $K(u)$, defined in chapter 1, equation (1), is the weighting function $w(x_i, t, h)$ which controls the probability of an observation x_i being censored in the semi-parametric procedure. We define

$$c = \int_u K(u)du,$$

$$K_1 = \frac{1}{c^2}\int_u K(u)^2 du$$

and

$$k_2 = \frac{1}{c}\int_u u^2 K(u)du.$$

Using small h expansions similar to used to approximate the MSE of an ordinary kernel estimate in Silverman (1986), section 3.3, we find that the two components of our local likelihood function have the following asymptotic values. If we define $u = (x - t)h^{-1}$, then

$$K\left(\frac{x-t}{h}\right)\log f(x,\theta) \simeq$$

$$K(u)\left(\log f(t,\theta) + hu\frac{d}{dt}\log f(t,\theta) + \frac{1}{2}h^2u^2\frac{d^2}{dt^2}\log f(t,\theta)\right)$$

and

$$\log\left(1 - c\int_x \frac{1}{c}K\left(\frac{x-t}{h}\right)f(x,\theta)dx\right) \simeq -chf(t,\theta)-$$

$$\frac{1}{2}c^2h^2f(t,\theta)^2 - \frac{1}{2}ch^3k_2\frac{d^2}{dt^2}f(t,\theta) - \frac{1}{3}c^3h^3f(t,\theta)^3.$$

When h is small, the partial derivative of the local likelihood function with respect to θ_i, the ith component of θ, can be approximated as

$$\frac{d}{d\theta_i} L_w(t, x, \theta, h) \simeq$$

$$cnh \left((\hat{g}(t) - f(t,\theta)) \left(\frac{d}{d\theta_i} \log f(t,\theta) \right) + ch \left(\frac{d}{d\theta_i} f(t,\theta) \right) (\hat{g}(t) - f(t,\theta)) + \right.$$
$$h^2 \left(S^* \left(\frac{d}{dt} \frac{d}{d\theta_i} \log f(t,\theta) \right) - c^2 f(t,\theta)^2 \left(\frac{d}{d\theta_i} f(t,\theta) \right) - \frac{1}{2} k_2 \left(\frac{d^2}{dt^2} \frac{d}{d\theta_i} f(t,\theta) \right) \right.$$
$$\left. \left. + \; c^2 \hat{g}(t) f(t,\theta) \left(\frac{d}{d\theta_i} f(t,\theta) \right) + \frac{1}{2} T^* \left(\frac{d^2}{dt^2} \frac{d}{d\theta_i} \log f(t,\theta) \right) \right) \right), \qquad (2)$$

where

$$\hat{g}(t) = \frac{1}{nhc} \sum_{i=1}^{n} K \left(\frac{x_i - t}{h} \right)$$

is the ordinary kernel density estimate of $g(t)$, using h as its bandwidth,

$$S^* = \frac{1}{nch^3} \sum_{i=1}^{n} (x_i - t) K \left(\frac{x_i - t}{h} \right)$$

and

$$T^* = \frac{1}{nch^3} \sum_{i=1}^{n} (x_i - t)^2 K \left(\frac{x_i - t}{h} \right).$$

The expectations of S^* and T^* are $g'(t)k_2 + O(h^2)$ and $g(t)k_2 + O(h^2)$ respectively.

As we would expect, if $f = g$, which implies that our model for the data is correct, then the expectation of equation (2) is 0. This can be shown by substituting in $g(t)$ for $f(t,\theta)$ in (2), and calculating the small h approximation to the expectation of $\hat{g}(t)$, up to the order of h^2. The latter, given in Silverman (1986), section 3.3, is

$$E(\hat{g}(t)) = g(t) + \frac{1}{2}h^2 k_2 \frac{d^2}{dt^2} g(t) + O(h^4). \tag{3}$$

Using equation (2), we can now show that when bandwidth h is small, our semi-parametric estimate differs from $\hat{g}(t)$ only by a term of order h^2.

To do this, we choose θ at any particular t such that $f(t,\theta)$ matches the non-parametric estimate at that point. Define this value of θ as $\theta_t{}^*$, where

$$f(t, \theta_t{}^*) = \hat{g}(t). \tag{4}$$

Then

$$f(t,\tilde{\theta}_t) = \hat{g}(t) + h^2 \hat{g}(t) \left(\frac{d}{d\theta_i} |_{\theta=\theta_t{}^*} f(t,\theta) \right)^{-1} \left(S^* \frac{d}{dt} \frac{d}{d\theta_i} |_{\theta=\theta_t{}^*} \log f(t,\theta) + \right.$$
$$\left. \frac{1}{2} T^* \frac{d^2}{dt^2} \frac{d}{d\theta_i} |_{\theta=\theta_t{}^*} \log f(t,\theta) - \frac{1}{2} k_2 \frac{d^2}{dt^2} \frac{d}{d\theta_i} |_{\theta=\theta_t{}^*} f(t,\theta) \right) + O(h^3). \tag{5}$$

When θ is a scalar, $\theta_t{}^*$ is simply found from equation (4), assuming such a value of θ exists. When θ is a vector, then the estimate $f(t,\tilde{\theta}_t)$ is found from solving the simultaneous equations produced from (2). To solve these, we must choose $\theta_t{}^*$ such that the multiple of h^2 in (5) is the same for all values of i.

Using equation (3) and the expectations of S^* and T^* given above, it is simple to calculate the expectation of equation (5), the small h approximation to the semi-parametric estimate. (**From now on, all derivatives (of any function) with respect to t will be written using 'prime' notation; i.e.** $\frac{d}{dt}g(t) = g'(t)$, $\frac{d^2}{dt^2}g(t) = g''(t)$, **etc.**) For small h,

$$E(f(t,\tilde{\theta}_t)) = g(t) + \frac{1}{2}h^2k_2\,(g''(t) - f''(t,\theta_t^*) + 2\beta_i(t)) + O(h^3), \qquad (6)$$

where

$$\beta_i(t) = \left(\frac{\frac{d}{d\theta_i}|_{\theta_i=\theta_i^*} f'(t,\theta)}{\frac{d}{d\theta_i}|_{\theta_i=\theta_i^*} f(t,\theta)} - \frac{f'(t,\theta_t^*)}{g(t)} \right) (g'(t) - f'(t,\theta_t^*)). \qquad (7)$$

Function $\beta_i(t)$ is equal to zero for all i when θ is of vector form, such as when $f \sim$ Normal, as opposed to when θ is of scalar form (for example, when $f \sim$ exponential). This is because θ_t^* is chosen such that the bias term in (6) is identical for all i. Since $(\frac{d}{d\theta_i}f(t,\theta_t^*))^{-1}\frac{d}{d\theta_i}f'(t,\theta_t^*)$ will take different values for different component vectors θ_i, in order for the bias term to be constant we must have the asymptotic equality

$$f'(t,\theta_t^*) = g'(t). \qquad (8)$$

Thus when θ is a vector, our choice of θ_t^* in equation (5) and in the following equations is simply the value which satisfies (4) and (8). (Plot 4a illustrates the choice of parameter $\theta = \theta_t^*$ when θ is a vector. Our chosen parametric family here is $f \sim$ Normal and we consider the limiting case, where as $n \rightarrow \infty$ and $h \rightarrow 0$, we take $\hat{g}(t) \simeq g(t)$.)

As our small h approximation to the semi-parametric estimate differs from $\hat{g}(t)$ only in terms of order h^2 and above, their asymptotic variances will be approximately the same. We use the variance approximation for $\hat{g}(t)$ given in Silverman (1986), section 3.3, such that for small h and large n,

$$Var(f(t,\tilde{\theta}_t)) = \frac{g(t)K_1}{nh}.$$

It now follows that the asymptotic MSE of the non-adaptive semi-parametric estimate can be written as

$$MSE(f(t,\tilde{\theta}_t)) = \frac{g(t)K_1}{nh} + \frac{1}{4}h^4 {k_2}^2 (g''(t) - f''(t,\theta_t^*) + 2\beta_i(t))^2, \qquad (9)$$

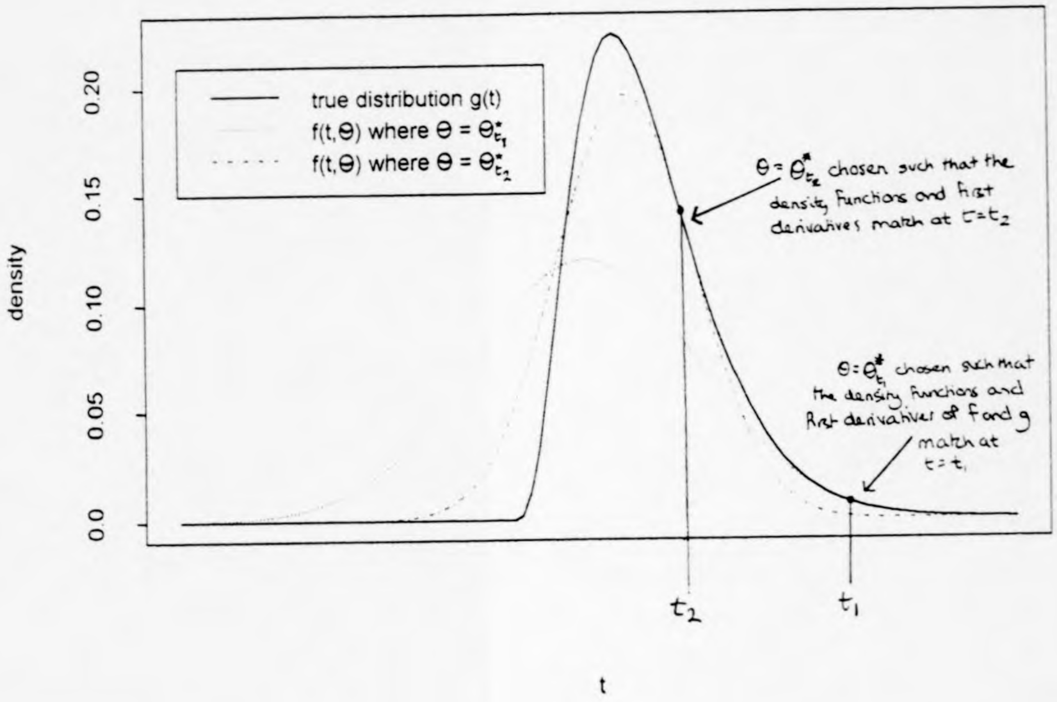
which simplifies to

$$MSE(f(t,\tilde{\theta}_t)) = \frac{g(t)K_1}{nh} + \frac{1}{4}h^4 {k_2}^2 (g''(t) - f''(t,\theta_t^*))^2$$

when θ is a vector.

The logical next step would be to choose h to minimise (9). However it is impractical to directly use the above theory to choose h, since we must first calculate θ_t^*. This initial step requires h itself in order to construct $\hat{g}(t)$!

PLOT 4a: Selecting parameter Θ_t^*



Sidestepping this problem, we consider equation (3). We expect our nonparametric kernel estimate to differ from the true distribution by a term of order h^2. Obviously we cannot attempt to choose θ_t^* by selecting it such that $f(t, \theta_t^*) = g(t)$ (and $f'(t, \theta_t^*) = g'(t)$ in the vector case), as the true density function $g(t)$ is unknown. Instead we replace $\hat{g}(t)$ and the first derivative of the true distribution in (4) and (8) by further 'preliminary' kernel estimates $\hat{g}^*(t)$ and $\hat{g}^{*\prime}(t)$.

As the following subsection will outline, we will choose a bandwidth h^* used to smooth these kernel functions under the same large n and small h assumptions as before, attempting to maximise the accuracy of these estimates to the true distribution. Equation (3) shows that we expect both $\hat{g}(t)$ and $\hat{g}^*(t)$ to differ from $g(t)$ only in terms of the of order the squares of their respective bandwidths h and h^*. Since we've assumed that h is small, and that n is large implying that our choice of h^* will be small too, then the difference between h^2 and h^{*2} will be small and preliminary kernel estimate $\hat{g}^*(t)$ should be close to $\hat{g}(t)$. Similarly we expect our estimate of $g'(t)$ to differ from the true value by a term of order h^2. Subsection 4.2.2 will consider kernel derivative estimation in more detail.

So in the practical application of this method we choose $\theta = \theta_t^*$ such that

$$f(t, \theta_t^*) = \hat{g}^*(t) \tag{10}$$

in the scalar case, or if a solution to (10) does not exist, as the minimiser of

$$(f(t, \theta_t^*) - \hat{g}^*(t))^2.$$

When θ is a vector, we choose $\theta = \theta_t^*$ such that

$$f(t, \theta_t^*) = \hat{g}^*(t)$$

and

$$f'(t, \theta_t^*) = \hat{g}^{*\prime}(t). \tag{11}$$

We then return to equation (9) as an estimate of the MSE of $f(t, \hat{\theta}_t)$.

The construction of this preliminary estimate is outlined in the next subsection. Despite the reliance of all of the above theory on small h and large n, this method of choosing h has performed satisfactorily for smaller samples, and has selected large values of h when this has been appropriate. Several of the examples given in chapters 5, 6 and 7 illustrate this.

4.2.2 Finding a preliminary estimate of true distribution g

Ordinary kernel estimates of $g(t)$, $g'(t)$ and $g''(t)$ are now required, initially to evaluate θ_t^*. We also need them as estimates of the true density function $g(t)$

and its derivatives $g'(t)$ and $g''(t)$, for use in (9) and in subsequent theory on choosing α in chapter 5. A reliable procedure for constructing them is therefore vital to these suggested methods for selecting the best values of h and α. We now require a bandwidth h^* for use in evaluating these preliminary kernel estimates.

Silverman (1986) gives several methods for choosing h^*. Most of these are designed to minimise the loss in accuracy of the density estimate $\hat{g}^*(t)$ to $g(t)$. Define $K^*(u)$ as the kernel function used in our kernel estimation of $g(t)$ and its derivatives. $K^*(u)$ may or may not be the same kernel function as $K(u)$, which is used as the weighting function in the censoring process driving the semi-parametric method. Then

$$c^* = \int_u K^*(u)du,$$

$$K_1{}^* = \frac{1}{c^{*2}} \int_u K^*(u)^2 du$$

and

$$k_2{}^* = \frac{1}{c^*} \int_u u^2 K^*(u) du$$

Silverman proves that the choice of h^* which minimises the MISE of the kernel estimate of $g(t)$ is approximately

$$k_2{}^{*-\frac{2}{5}} K_1{}^{*\frac{1}{5}} \left(\int g''(t)^2 dt \right)^{-\frac{1}{5}} n^{-\frac{1}{5}}, \tag{12}$$

where h^* is initially assumed small and the sample size n is large.

In this case we are interested not only in $g(t)$ but its first and second derivatives as well. So for example, while h^* chosen to minimise the MISE of

$$\hat{g}^*(t) = \frac{1}{nh^*c^*}\sum_{i=1}^{n} K\left(\frac{x_i - t}{h^*}\right)$$

will produce a good estimate of $g(t)$, it may not lead to accurate estimates

$$\hat{g}^{*\prime}(t) = \frac{d}{dt}\frac{1}{nh^*c^*}\sum_{i=1}^{n} K\left(\frac{x_i - t}{h^*}\right)$$

and

$$\hat{g}^{*\prime\prime}(t) = \frac{d^2}{dt^2}\frac{1}{nh^*c^*}\sum_{i=1}^{n} K\left(\frac{x_i - t}{h^*}\right)$$

of $g'(t)$ and $g''(t)$ respectively. This problem is likely to occur if h^* is too small and errs towards undersmoothing our density estimate; then at any point t, the kernel method will give us $\hat{g}(t) \simeq g(t)$ but any noise in our kernel density estimate around t will mean that we obtain very poor estimates of the first and second derivatives both in terms of magnitude and sign. Therefore we need a procedure for selecting h^* with respect to the expected accuracy of $\hat{g}'(t)$ and especially the extremely volatile $\hat{g}''(t)$.

A simple formula for choosing a best h^* can be obtained by minimising the MISE of our estimate of the second derivative. Several combinations of this along with the formulae for best h^* resulting from minimising the MISE of our estimate of the first derivative and of $g(t)$ itself were also tried, but involved more calculation and did not perform as well, occasionally choosing h^* too small. Bandwidth h^* selected by this method will be slightly larger than that given by equation (12), thus avoiding any of the noise in our estimate of $g(t)$ which causes poor estimates of $g''(t)$. Such oversmoothing produces an estimate of $g(t)$ which is marginally less accurate than that obtained when using the bandwidth given in equation (12), especially in regions of high density, but does ensure that the potentially larger inaccuracies when estimating the first and especially the second derivative do not occur.

Now assume that $K^*(u)$ is the Gaussian kernel function such that

$$K^*(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right).$$

(This assumption will cause difficulties when the true distribution is clearly non-Gaussian in the tails, for example in the $g \sim$ exponential case. Such situations are dealt with later). If $K^*(u)$ is the Gaussian kernel, then using small h^* approximations given in Silverman (1986) and Wand and Jones (1995), we find that

$$MISE(\hat{g^*}''(t)) = \int_t \left(E(\hat{g^*}''(t)) - g''(t)\right)^2 dt + \int_t var\left(\hat{g^*}''(t)\right) dt$$

$$\simeq \frac{h^4}{4}\int_t (g^{iv}(t))^2 dt + \frac{3}{8\sqrt{\pi} n h^5}.$$

To minimise this quantity we choose

$$h^* = n^{-\frac{1}{9}} \left(\int_t (g^{iv}(t))^2 dt\right)^{-\frac{1}{9}} \left(\frac{15}{8\sqrt{\pi}}\right)^{\frac{1}{9}}. \tag{13}$$

Following the policy of Silverman (1986), we now estimate $g^{iv}(t)$ by replacing g with a suitable parametric family whose density function we believe to be similar in shape to $g(t)$. The logical next-step in this setting is to use the fourth derivative of our parametric 'guess' $f(t,\theta)$ to estimate $g^{iv}(t)$, and insert it into equation (13). We use the MLE $\hat{\theta}$ to estimate θ where necessary. For example, if f is a Normal distribution, with variance σ^2 estimated by $\hat{\sigma}^2$, where $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)^T$, then

$$\int_t (f^{iv}(t,\hat{\theta}))^2 dt = \frac{105}{32\sqrt{\pi}\hat{\sigma}^9},$$

and so we choose

$$h^* = n^{-\frac{1}{9}} \hat{\sigma} \left(\frac{12}{21}\right)^{\frac{1}{9}}.$$

It is also recommended (as in Silverman (1986)) that since h^* is dependent on the variability of the data, that a more robust measure of spread than the sample standard deviation is used, such as

$$A^* = \min(\text{sample s.d } \hat{\sigma},\ \text{interquartile range}/1.34).$$

For some choices of f, it may be impossible to find

$$\int_t (f^{iv}(t,\hat{\theta}))^2 dt$$

analytically. Since our choice of h^* is based on the shape and variance of the distribution we seek to estimate, our best option in these case is to estimate it by

$$\int_t (p^{iv}(t))^2 dt,$$

where $p(t)$ is the probability density function of similar shape to $f(t,\theta)$ and $g(t)$. For example, when our chosen parametric family f is a Gamma or a Weibull distribution, with fitted parameters indicating a shape roughly similar to that of a Normal distribution, then choosing $p \sim$ Normal has proved a satisfactory solution. In fact, **in all cases where a histogram of the data has appeared unimodal or multimodal, without suggesting that $g(t)$ has a bounded domain**, taking

$$h^* = n^{-\frac{1}{9}} A^* \left(\frac{12}{21}\right)^{\frac{1}{9}} \qquad (14)$$

has led to adequate estimates of the derivatives of $g(t)$.

Therefore I recommend using formula (14) as the bandwidth for the preliminary kernel estimate in all cases. Though it may oversmooth, especially when the data appears multimodal, the fine accuracy of our estimate is of limited importance. The essential requirement is to avoid large random errors. If we choose h^* from formula (14), it is simple to calculate and provides sensible rough estimates of $g(t)$, $g'(t)$, and $g''(t)$ for selecting $\theta_t{}^*$, evaluating (9), and for further use in later chapters.

To handle cases where the true distribution is drastically different in shape from a Normal distribution we require a slight adjustment to our approach. Data sets which appear to come from an exponential distribution are the most common example of this. Problems in producing preliminary estimates occur due to the combination of a discontinuity at $t = 0$ and the bounded interval of t on which the density function is defined. If we have taken $f \sim$ exponential, then neither using a Gaussian kernel, nor replacing $f(t,\theta)$ by $p(t)$, where $p(t)$ is the Normal density function, is appropriate. However, rather than advocating a totally different method for these cases, a more sensible solution is to use reflected kernel density estimation. Instead of considering the data set $X = (x_1, ..., x_n)$, we augment it by its reflection in the y axis giving a data set $X^* = (-x_1, ..., -x_n, x_1, ..., x_n)$. We now use the above methods to evaluate estimates $\bar{g}(t)$, $\bar{g}'(t)$ and $\bar{g}''(t)$ of the density function of the true distribution of X^*, and its derivatives. Using these methods is now justified since the true distribution of X^* is of unimodal shape roughly similar to that of a Normal distribution. To get our final estimates $\hat{g}^*(t)$, $\hat{g}^{*\prime}(t)$ and $\hat{g}^{*\prime\prime}(t)$, we consider $\bar{g}$ only·to the right of the y axis and multiply by 2, so that for the nth derivative,

$$\hat{g}^{*n}(t) = 2\bar{g}^n(t) \qquad \forall t \geq 0,$$

and

$$\hat{g}^{*n}(t) = 0 \qquad \forall t < 0. \tag{15}$$

See Silverman (1986), page 30 for further discussion of this method.

This worked passably well when tried on several data sets with density functions sharing the bounded domain, discontinuity and extreme non-Normality of the exponential distribution. When selecting an overall bandwidth for our adaptive semi-parametric method we are only interested in estimating behaviour around the boundary point $t = t^*$, which is located at the sample mean for the $f \sim$ exponential case. The reflection method suggested above gives good estimates of $g(t)$ and its derivatives in this region, enabling the suggested method of selecting h to perform satisfactorily.

But in several of the methods of selecting α to be introduced in chapter 5, estimates of the true density and its first two derivatives are required for all t. The reflected kernel technique gives poor results around $t = 0$, where due to the structure of (15), our estimate is flat. Therefore these methods of selecting α perform poorly for distributions which we believe to be similar in shape to an exponential. However, given the general awkwardness of coping with bounded distributions, these difficulties and inaccuracies have to be suffered, since several other solutions for choosing h^* and estimating $g(t)$, $g'(t)$ and $g''(t)$ that I investigated do no better, and lack the simplicity of this idea. As a whole, our semi-parametric method deals very well with data which come from bounded and discontinuous distributions. For example, if we believe the true distribution to be exponential, we can fit this distribution to the data, which we cannot do in non-parametric estimation. Errors in these preliminary rough estimates of $g(t)$ and its derivatives can largely be tolerated since they are not at the sharp end of the actual semi-parametric density estimation process; they are just being applied in suggested methods for selecting h and α.

Kernel density estimation using h^* as given in equation (14), with the adjustment described applied when the domain of f is bounded, is simple to use and gave us sufficiently accurate estimates of $g(t)$, $g'(t)$ and $g''(t)$ for use in selecting an overall bandwidth for our adaptive semi-parametric method.

We now employ these estimates to do just that.

4.2.3 An automatic formula for 'best' h

Differentiating equation (9) with respect to h, setting equal to 0 and solving, leads us to our best choice of h for use in our adaptive semi-parametric method in terms of minimising the MSE at the boundary point t^*. This is

$$h = h' = (g''(t^*) - f''(t^*, \theta_{t^*}{}^*) + 2\beta_i(t^*))^{-\frac{2}{5}} k_2{}^{-\frac{2}{5}} n^{-\frac{1}{5}} K_1{}^{\frac{1}{5}} g(t^*)^{\frac{1}{5}}. \qquad (16)$$

Appropriately our overall bandwidth is chosen with respect to the difference between f and g. An increase in this difference indicates that the parametric family is becoming less accurate to g; thus the resultant decrease in the value of h chosen is desirable, since we would want a more non-parametric estimate. When θ is a scalar, our choice of h is dependent on the differences between the first derivatives (within $\beta_i(t)$) and between the second derivatives of the density functions defining the true and parametric distributions. If θ is a vector, it is dependent only on the differences between the second derivatives. The selection of $\theta = \theta_t{}^*$, the parameter value at which the density function of our chosen parametric family is evaluated in (16), is explained in section 4.2 and illustrated in plot 4a.

Rather like the formula (12) for the bandwidth for an ordinary kernel estimate of $g(t)$, equation (16) contains the unknown true probability density function $g(t)$ and its derivatives. However we need only to choose h to be suitable at one point $t = t^*$ rather than to find a value which performs acceptably over a whole range of t; using adaption should ensure that. As outlined in subsection 4.2.2, we can obtain sufficiently accurate point estimates of the true density and its first and second derivatives using a slightly oversmoothed ordinary kernel density estimate.

In cases where g is extremely non-Normal, such as when it resembles an exponential distribution, formula (16) has occasionally produced values of h which were too small. An adjustment of this automatic method, to be introduced in Chapter 6, which facilitates the choosing of a larger overall bandwidth with respect to an index of prior belief about g, is recommended in these or any other circumstances which have resulted in too small a value of h being produced.

As our sample size $n \to \infty$, our optimal choice of h will move towards zero at a very slow rate. The MSE of the semi-parametric method at t^*, given the optimal choice of $h = h'$, is equal to

$$\frac{5}{4}k_2{}^{\frac{2}{5}}K_1^{\frac{4}{5}}g(t^*)^{\frac{4}{5}}(g''(t^*) - f''(t^*, \theta_{t^*}{}^*) + 2\beta_i(t^*))^{\frac{2}{5}}n^{-\frac{4}{5}},$$

which is a monotone decreasing function of n. Because the application of adaption obviously affects this limiting behaviour of our estimate elsewhere in its range, it is considered in more detail for the different methods of choosing the amount of adaption α, which are explored in chapter 5.

4.3 When more than one boundary point exists

For many distributions f, there is more than one boundary point because there are several values of t such that $f(t, \hat{\theta}) = \lambda$. For example, if f is a Normal distribution there will be two boundary points located at one sample standard deviation either side of the sample mean. If given $f \sim N[\mu, \sigma^2]$, it is clear that

$$\begin{aligned}\lambda &= \exp\left(\int_t f(t, \hat{\theta}) \log f(t, \hat{\theta}) dt\right) \\ &= \frac{1}{\sqrt{2\pi}\hat{\sigma}} \exp(-\frac{1}{2}).\end{aligned}$$

Thus when $f(t, \hat{\theta}) = \lambda$, then

$$\left(\frac{t - \hat{\mu}}{\hat{\sigma}}\right)^2 = 1$$

implying that

$$t = \hat{\mu} \pm \hat{\sigma}.$$

Our suggested 'automatic' choice of h is made by minimising the MSE of the semi-parametric method at a boundary point. In the case of multiple boundary points, a method that has given good results is to select the value of h which minimises the MSE at the ' Maximum Difference Boundary Point', denoted MBP. This point has the property that, given that there exist m other boundary points $b_1, ..., b_m$, then

$$(g(MBP) - f(MBP, \theta))^2 \geq (g(b_i) - f(b_i, \theta))^2$$

where $1 \leq i \leq m$. In other words, the MBP is the boundary point at which there is the greatest difference between the true distribution $g(t)$ (which we estimate using kernel density estimation as described in subsection 4.2.2) and our parametric guess $f(t, \theta)$ (with θ estimated by MLE $\hat{\theta}$).

While this may appear a somewhat ad-hoc procedure, it has some useful properties in addition to being relatively straightforward to implement. Minimising the MSE at the boundary point at which the difference in $f(t, \theta)$ and $g(t)$ is the greatest will require the choice of a smaller h, which has two advantages.

Firstly it conforms with our original assumption of small h when using the approximation for the MSE. Except for the occasional case where the extreme non-Normality of the true density function posed problems in the preliminary estimation stage described in subsection 4.2.2, our values of h chosen in this way have still been large enough to ensure smooth tails for our semi-parametric estimate. In fact, in a large series of trials, only when we chose $f \sim$ exponential did our automatically selected bandwidth ever lead to an inadequately smoothed adaptive semi-parametric density estimate.

Secondly, consider the case where $f(t,\theta)$ is a very poor estimate of $g(t)$ at a boundary point (at which adaption has no effect). Erring on the side of small h here will at least reduce the influence of parametric family f and instead provide a more non-parametric estimate. This should give a more accurate reflection of the amount of data in the region, and run smoothly into the even more non-parametric estimate produced in the neighbouring region where adaption is making $h_{t,\alpha} \leq h$.

However, if we have a larger overall bandwidth we may encounter problems at the boundary points. If $f(t,\theta)$ differs dramatically from $g(t)$ here, then our largely parametric estimate will be poor. As well as this, we will need to choose a large value of α to get a more accurate largely non-parametric estimate in the regions of high density, where $h_{t,\alpha} \leq h$. But if α is too large, we will move rapidly from a region where $h_{t,\alpha}$ is very small towards the boundary point, where $h_{t,\alpha} = h$ is large. An awkwardly sharp change in the size and shape of our density estimate $f(t, \hat{\theta}_{t,\alpha})$ can occur, as it moves between approximating $\hat{g}_L(t)$ and approximating $f(t, \hat{\theta})$ over a very small interval of t.

A smaller choice of h means that we require less adaption to get small enough values of $h_{t,\alpha}$ to provide the best estimate in areas where $f(t, \theta)$ and $g(t)$ are large. How exactly to choose the right amount of adaption is discussed in the next section. A further brief summary of this chapter in the light of the ideas developed in the next is given at the start of section 5.4.

5 Several automatic methods of choosing α

5.1 Defining a region of interest

The following methods for selecting α will attempt to maximise the accuracy of our adaptive semi-parametric method only over the region D^*, defined in chapter 3 as

$$D^* = (t : h_{t,\alpha} \leq h). \tag{1}$$

This restriction is both convenient and justifiable.. When choosing α it is necessary to use either small or large $h_{t,\alpha}$ approximations to the adaptive semi-parametric method, since no exact expansion for $f(t, \hat{\theta}_{t,\alpha})$ exists. If α is greater than zero, $h_{t,\alpha}$ could vary dramatically in size over t, so we cannot use a single approximation for $f(t, \hat{\theta}_{t,\alpha})$ for all t. Either we attempt to combine approximations to cover the whole range of t, or we concentrate on one of the two distinct regions formed when α takes a positive value, which are D^*, and its complement D^{*C} where $h_{t,\alpha} \geq h$.

The region D^* is where the greatest scope exists for improving the accuracy of our estimate by applying the right amount of adaption. We allow α to 'take care' of the distinct region D^{*C}, where applying adaption will see our adaptive semi-parametric estimate tend towards a smooth parametric estimate, with any resultant loss in accuracy being both small and bounded.

In the following methods of selecting an optimal value of α, **we will ignore D^{*C} and concentrate solely on D^***, enabling us to use small $h_{t,\alpha}$ approximations to our adaptive semi-parametric estimate. We assume that h and α are always chosen large enough such that the tails of $f(t,\hat{\theta}_{t,\alpha})$ are sufficiently smoothed. The possibility of this not being the case is one motivating factor behind the ideas of chapter 6. These suggest different methods for choosing the overall bandwidth, which ensure that it is large enough to avoid a noisy density estimate in D^{*C}. In terms of minimising the loss in accuracy of $f(t,\hat{\theta}_{t,\alpha})$ to $g(t)$ over all t, the choices of α given by the following methods may well be sub-optimal since we are ignoring a subset of the range of t, but their practical use is demonstrated in section 5.3.

There is a final advantage of working on D^* alone. In all of the following methods integration will be necessary, and when the integral has no simple analytical solution, numerical methods such as Simpson's Rule will have to be applied. The accuracy and ease of these calculations is enhanced if we are integrating over a finite interval. By definition, D^* must always be bounded, since the probability density function $f(t,\hat{\theta})$ which determines its location must either be defined only on a bounded domain, or decrease to 0 as $|t| \rightarrow \infty$. If the latter happens, as $|t| \rightarrow \infty$ we will eventually pass the 'last' boundary point before infinity and enter D^{*C}.

5.2 Four automatic selections of α

The design of the automatic method for choosing the overall bandwidth introduced in section 4.2 produces values of h which are likely to err on the small side of optimal. It makes sense to use the following methods for choosing α in conjunction with an automatically chosen bandwidth h. Then the small $h_{t,\alpha}$ approximations which underpin the following procedures should hold fairly well for the region D^*. By definition

$$h_{t,\alpha} \leq h \qquad \forall t \in D^*,$$

thus if h is small, $h_{t,\alpha}$ will be.

Problems do occur when we want to use a larger (handpicked) value of h than that chosen automatically. At values of t in D^* near to the boundary point(s), $\Upsilon(t)^{-\alpha}$ is only marginally less than one, and so $h_{t,\alpha} \simeq h$. If h is large, then in these areas the inaccuracy of our approximations will affect our 'best' choice of α. This difficulty is examined in more detail for each of the methods and examples.

Under several of the following procedures for selecting α, the choices of h and α can be made in any order. I will show that when both h and α are chosen automatically, our selection of α is independent of the sample size, and the type of kernel used in the censoring process which is at the heart of the semi-parametric method.

5.2.1 Method (i): Approximating to a local kernel density estimate

Copas (1995a) states that when h is very small, the semi-parametric estimate at t approximates the ordinary kernel estimate with bandwidth h at that point. To see this, consider the small h approximation of chapter 4, equation (5), and ignore all terms of order h^2 and above. If we are varying our bandwidth with t, then our adaptive semi-parametric estimate approximates to the local kernel density estimate, where the local bandwidth at any point t is $h_{t,\alpha}$.

Therefore, as an automatic method of selecting α, we could choose $\alpha = \alpha'$ where α' is the best choice of adaption parameter for the local kernel method over region D^*. Previous work on kernel density estimation with a variable bandwidth has been concerned with finding an optimal value of α when considering estimation of $g(t)$ over the full range of t. This has made any small or large $h_{t,\alpha}$ approximations impossible. As referred to in section 1.4, choosing $\alpha = 0.5$ for the slightly different varying kernel method has useful bias reduction properties, but a fixed value of α for all cases would be inappropriate here. Instead we use the fact that $h_{t,\alpha}$ will always be small when $t \in D^*$, and extend the small h approximation to the MISE of an ordinary kernel density estimate to the local case. We then choose the value of α which minimises this loss function over D^*.

The equations for the bias and the variance of the ordinary kernel density estimate translate automatically to the local case, so we can say that for any t such that $h_{t,\alpha} < h$, if we approximate our adaptive semi-parametric estimate $f(t, \tilde{\theta}_{t,\alpha})$ by the local kernel density estimate with small $h_{t,\alpha}$ and assume a large sample size n, then

$$bias(f(t,\tilde{\theta}_{t,\alpha})) = E_g(f(t,\tilde{\theta}_{t,\alpha})) - g(t) = \frac{1}{2}{h_{t,\alpha}}^2 g''(t)k_2 + O(h^3)$$

and

$$var(f(t,\tilde{\theta}_{t,\alpha})) \simeq n^{-1}{h_{t,\alpha}}^{-1}g(t)K_1. \tag{2}$$

The MSE is the sum of the squared bias and the variance of the estimate.

Writing $h_{t,\alpha}$ as in equation (1) of chapter 4 and estimating θ by its MLE $\hat{\theta}$, the MISE of the adaptive semi-parametric method over D^* can be approximated as

$$MISE_{D^*}(f(t,\hat{\theta}_{t,\alpha})) = \frac{1}{4}h^4 {k_2}^2 \int_{D^*} \Upsilon(t)^{-4\alpha} g''(t)^2 dt$$
$$+ n^{-1} h^{-1} K_1 \int_{D^*} \Upsilon(t)^{\alpha} g(t) dt. \qquad (3)$$

In the above equations h is our chosen overall bandwidth. The definition of $\Upsilon(t)$, a function of t, is given in equation (1) of chapter 4. Constants K_1 and k_2 were defined in subsection 4.2.1. We now find the value of α that minimises the above equation. Differentiating with respect to α and setting equal to zero, we see that our best choice of α, which minimises (3), will solve the equation

$$h^5 {k_2}^2 n K_1^{-1} \int_{D^*} \Upsilon(t)^{-4\alpha} \log \Upsilon(t) g''(t)^2 dt = \int_{D^*} \Upsilon(t)^{\alpha} \log \Upsilon(t) g(t) dt. \qquad (4)$$

I found this easy to solve using a simple minimisation programme on a computer. Since D^* will always be bounded, numerical approximations to the integrals are straightforward to calculate. Also required are ordinary kernel density estimates for $g(t)$ and its second derivative, the evaluation of which are described in subsection 4.2.2.

Inserting formula (16) from chapter 4 for the best overall bandwidth h into the above equation, we see that when h is chosen by the automatic method of chapter 4, our automatic choice of α using method (i) satisfies

$$\int_{D^*} \Big(g''(t)^2(g''(t^*) - f''(t^*, \theta^*{}_{t^*}) + 2\beta_i(t^*))^{-2} g(t^*)\Upsilon(t)^{-4\alpha}$$

$$-g(t)\Upsilon(t)^{\alpha}\Big) \log \Upsilon(t) dt = 0, \tag{5}$$

and is thus independent of both sample size n and the type of scaled kernel function $K(u)$ used to perform the weighting in the local likelihood function. This convenient property does not hold for general h, giving us another reason to use the automatic method of selecting h.

When h is selected in this way, then in the vector case where $\beta_i(t) = 0$, α will be dependent on the difference between the second derivatives of $f(t, \theta_t^*)$ and $\hat{g}(t)$ at t^*. The larger this difference is, the smaller the optimal α value will be; a desirable property, since it reduces the possibility of roughness in our density estimate at the boundary points. This is likely to occur when α is large and the behaviour of our model based upon parametric family f around a boundary point differs dramatically from that of true distribution g.

There are two problems with this method, namely the need to estimate

$g''(t)$ and, when h gets large, the aforementioned poor approximation of $f(t, \hat{\theta}_{t,\alpha})$ by $\hat{g}_L(t)$ near to the boundary points.

We can consider the first to be a necessary evil, and it can be tolerated as long as we err towards oversmoothing our preliminary kernel estimate $\hat{g}^*(t)$, for reasons given in subsection 4.2.2.

However the poor quality of the approximation underpinning method (i) when h is large is a more serious handicap. Practical tests of this method for a variety of different distributions f and g have shown that it produces good choices of α when the overall bandwidth h is small. In this case $h_{t,\alpha}$ will also be small for all $t \in D^*$, and the approximation we are using is reasonable. As h increases, both this approximation and our choice of α become less satisfactory, with the latter increasing at too fast a rate. Even for overall bandwidths of around 1 or 2 sample standard deviations, method (i) chooses an α value so large that $h_{t,\alpha}$ is very small in regions of high density, leading to a spiky, uneven density estimate in much of D^*. As the examples in section 5.3 demonstrate, this will give a much larger non-approximated MISE of $f(t, \hat{\theta}_{t,\alpha})$ than a smaller value of α would, and is a poor estimate of the true density.

5.2.2 Method (ii): Least Squares Cross-Validation

Using the same range limitation and small $h_{t,\alpha}$ approximation as in method (i), we can sidestep the problem of having to estimate $g''(t)$ by using Least Squares Cross-Validation (LSCV). Silverman (1986) outlines the procedure for ordinary kernel density estimation and there is a simple analogy for the local kernel method over a restricted range.

As before we seek to minimise the MISE over D^*,

$$\int_{D^*}(g(t) - f(t, \bar{\theta}_{t,\alpha}))^2 dt,$$

with respect to α, where $f(t, \bar{\theta}_{t,\alpha})$ is approximated by the local kernel density estimate $\hat{g}_{L,\alpha}(t)$. That is, assuming $h_{t,\alpha}$ is small, we take

$$f(t, \bar{\theta}_{t,\alpha}) \simeq \hat{g}_{L,\alpha}(t) = h_{t,\alpha}{}^{-1} n^{-1} c^{-1} \sum_{i=1}^{n} K\left(\frac{t - x_i}{h_{t,\alpha}}\right), \qquad (6)$$

where c is as defined in subsection 4.2.1.

Since $g(t)$ is independent of α, the choice of α which minimises the MISE over the region D^* will also minimise

$$R(f) = R(f(t, \bar{\theta}_{t,\alpha})) = \int_{D^*} f(t, \bar{\theta}_{t,\alpha})^2 dt - 2\int_{D^*} f(t, \bar{\theta}_{t,\alpha}) g(t) dt. \qquad (7)$$

We can rewrite the first integral in (7) using our approximation (6). For the second half of $R(f)$ we use (6) again, enabling an initial approximation of

(7) by

$$\int_{D^*} \hat{g}_{L,\alpha}(t)^2 dt - 2\int_{D^*} \hat{g}_{L,\alpha}(t) g(t) dt.$$

We then consider the expectation of the second term using the following formulae.

Define $\hat{g}_{L,\alpha,(-i)}$ to be the local kernel density estimate of true density function $g(t)$ constructed from all of the n data points except for x_i. When estimating at t, this is written as

$$\hat{g}_{L,\alpha,(-i)}(t) = (n-1)^{-1} h_{t,\alpha}{}^{-1} c^{-1} \sum_{j \neq i} K\left(\frac{t - x_i}{h_{t,\alpha}}\right).$$

Assume there are l data points within region D^*, and that n is large, enabling us to approximate the probability of any observation being in D^* by ln^{-1}. Since the expectation of $\hat{g}_{L,\alpha}(t)$ depends only on the kernel function and not on the sample size, then

$$E\int_{D^*} \hat{g}_{L,\alpha}(t) g(t) dt = E\int_{D^*} \hat{g}_{L,\alpha,(-i)}(t) g(t) dt$$

$$= E\left(\hat{g}_{L,\alpha,(-i)}(x_i)/x_i \in D^*\right) \int_{D^*} g(t) dt \simeq \frac{l}{n} E\left(\hat{g}_{L,\alpha,(-i)}(x_i)/x_i \in D^*\right)$$

$$= \frac{l}{n} E\left(l^{-1} \sum_{i:x_i \in D^*} \hat{g}_{L,\alpha,(-i)}(x_i)\right) = E\left(n^{-1} \sum_{i:x_i \in D^*} \hat{g}_{L,\alpha,(-i)}(x_i)\right). \qquad (8)$$

So using this expectation to estimate the second term in equation (7), we choose α to minimise the score function

$$M_o(\alpha) = \int_{D^*} \hat{g}_{L,\alpha}(t)^2 dt - 2n^{-1} \sum_{i:x_i \in D^*} \hat{g}_{L,\alpha,(-i)}(x_i). \tag{9}$$

This score can be expressed in a more suitable form for computation, such as

$$\begin{aligned} M_o(\alpha) &= \int_{D^*} \hat{g}_{L,\alpha}(t)^2 dt - 2n^{-1} \sum_{i:x_i \in D^*} (n-1)^{-1} c^{-1} \sum_{j \neq i} h_{x_i,\alpha}{}^{-1} K\left(\frac{x_i - x_j}{h_{x_i,\alpha}}\right) \\ &= \int_{D^*} \hat{g}_{L,\alpha}(t)^2 dt - 2n^{-1} \sum_{i:x_i \in D^*} (n-1)^{-1} c^{-1} \sum_{j} h_{x_i,\alpha}^{-1} K\left(\frac{x_i - x_j}{h_{x_i,\alpha}}\right) \\ &\quad + 2n^{-1} \sum_{i:x_i \in D^*} (n-1)^{-1} c^{-1} h_{x_i,\alpha}{}^{-1} K(0). \end{aligned} \tag{10}$$

The first part of this sum can be calculated using numerical integration. We then found the value of α which minimised $M_o(\alpha)$ using a simple computer minimisation routine, though obtaining this for any particular example required slightly more computer time than when using method (i).

The advantages of using LSCV are that we no longer need to estimate the second derivative of the true distribution, or to use the small $h_{t,\alpha}$ approximation to the MSE of a local kernel estimate. Yet the results gained from using this method were very similar to those from method (i). Sensible applicable choices of α were achieved when the overall bandwidth h was small, but for larger values of h, the values of α chosen were too great. When used in the adaptive semi-parametric method, this gave a spiky estimate in region D^*, with sharp changes in gradient appearing around the boundary points in some cases. Despite the use of a large n approximation, the success of this method did not appear dependent on sample size, working well with small samples provided h was small.

Like method (i), method (ii) relies upon a small $h_{t,\alpha}$ approximation of the adaptive semi-parametric method in D^* by the local kernel estimate. It is the inaccuracy of this approximation as the overall bandwidth h increases which results in the poor performance of both these methods. Method (i) and method (ii) achieved good results when used in tandem with the automatic method of choosing h, and are worth considering in this context. They are also convenient for cases where f differs dramatically from the Normal distribution, since they rely less than the following methods on the preliminary derivative estimation of the true distribution for all t, outlined in subsection 4.2.2.

5.2.3 Method (iii): Using a direct small $h_{t,\alpha}$ approximation to the MSE of the adaptive semi-parametric method

The small $h_{t,\alpha}$ approximation used in method (i) is really a double approximation. We initially use the fact that our adaptive semi-parametric density estimate converges to a non-parametric local kernel estimate as $h_{t,\alpha}$ decreases to zero, and having assumed that $h_{t,\alpha}$ is small, we use a further approximation to the MSE of the local kernel density estimate. It is possible to cut out a stage by adapting the small h approximation to the MSE of the semi-parametric method from Copas (1995b), with a resultant gain in simplicity and accuracy. This approximation was introduced and derived in chapter 4, equations (2) to (9). Though it also relies on n being large, this method has still performed well with smaller sample sizes.

To find the MSE of the adaptive semi-parametric method at t, we replace h in chapter 4, equation (9), by $h_{t,\alpha}$, which is defined in chapter 4, equation (1). After integrating over D^*, the region in which a small $h_{t,\alpha}$ approximation will be valid, we choose α to minimise this integral, which is the approximate MISE over D^* of the adaptive semi-parametric method. It can be written as

$$MISE_{D^*}(f(t,\hat{\theta}_{t,\alpha})) = \frac{1}{4}h^4 {k_2}^2 \int_{D^*} \Upsilon(t)^{-4\alpha}(g''(t) - f''(t,{\theta_t}^*) + 2\beta_i(t))^2 dt$$
$$+ n^{-1}h^{-1}K_1 \int_{D^*} g(t)\Upsilon(t)^{\alpha} dt, \qquad (11)$$

where $f(t,{\theta_t}^*)$ and $\beta_i(t)$ are defined in section 4.2.

(This differs from the MISE of the local kernel estimate over D^* given in equation (3), which was used as an approximation to the MISE of $f(t,\hat{\theta}_{t,\alpha})$ in method (i), only in that $g''(t)$ is replaced by

$$g''(t) - f''(t,{\theta_t}^*) + 2\beta_i(t). \qquad (12)$$

Since this term is independent of α, the following results can be obtained by simply replacing $g''(t)$ in equations (4) and (5) by equation (2)).

Differentiating (11) with respect to α and setting equal to zero, we find that the best choice of α evaluated using method (iii), which aims to minimise (11), solves the equation

$$h^5 {k_2}^2 n K_1^{-1} \int_{D^*} \Upsilon(t)^{-4\alpha} \log \Upsilon(t)(g''(t) - f''(t,{\theta_t}^*) + 2\beta_i(t))^2 dt$$
$$= \int_{D^*} \Upsilon(t)^{\alpha} \log \Upsilon(t) g(t) dt. \qquad (13)$$

If we assume that the overall bandwidth is chosen by our automatic method, with the maximum difference boundary point being at $t = t^*$, then the best choice of α also satisfies

$$\int_{D^*} \Big((g''(t) - f''(t, \theta_t^*) + 2\beta_i(t))^2 (g''(t^*) - f''(t^*, \theta_{t^*}^*) + 2\beta_i(t^*))^{-2} g(t^*) \Upsilon(t)^{-4\alpha} \\ - \Upsilon(t)^{\alpha} g(t)\Big) \log \Upsilon(t) dt = 0. \qquad (14)$$

Both equations (13) and (14) were easy to solve using numerical integration and a computer minimisation package. We used ordinary kernel density estimation to estimate $g(t)$ and its derivatives as outlined in subsection 4.2.2. However, unlike in method (ii), we need estimates for $g(t)$, $g'(t)$ and $g''(t)$ over a range of target points, rather than just at a single boundary point. In cases where preliminary kernel estimates of these values are liable to be inaccurate at some values of t, such as when g appears to be an extremely non-Normal distribution such as the exponential, then this method is less appropriate. Method (iii) has the same properties as method (i) when used in conjunction with the automatically chosen h, namely that it is independent of the sample size n, and of the scaled kernel function $K(u)$ defined in chapter 1, equation (1), which performs the weighting within our local likelihood function.

This method again performs well for small values of our overall bandwidth. But while the choices of α for larger values of h were not as excessive as those from methods (i) and (ii), they still lead to overly small values of $h_{t,\alpha}$. It shares with method (i) the problem of estimating $g''(t)$ but this time not just a one point. While the small $h_{t,\alpha}$ approximation to the MISE of $f(t,\tilde{\theta}_{t,\alpha})$ is better than that used in (i) and (ii), it still limits our scope when we want to use a larger bandwidth.

5.2.4 Method (iv): Minimising the difference from $h_{t,\alpha}$ to the optimal local bandwidth at t

The ongoing problem of the small $h_{t,\alpha}$ approximation breaking down is finally solved by method (iv), using a combination of making the accuracy of this approximation less crucial to our result, and introducing a 'safety net'. Instead of attempting to minimise the difference between our estimate and the true distribution, we will design a method which chooses α to minimise the integrated distance between $h_{t,\alpha}$ and the 'best' choice of local bandwidth at each value of t. We work over the same interval D^* as before, though we rely on different approximations to those used in previous methods. At any point t, our optimal local bandwidth is defined as $h_{opt}(t)$, and is chosen with respect to minimising the small h approximation of the MSE of the semi-parametric method. We use exactly the same formulation as when automatically selecting our best overall bandwidth h, which is given in chapter 4, equation (16). However, here we apply it at any value of t within D^* rather than just at a single boundary point.

Consider the choice of α defined as the maximum of $(0, \alpha^*)$, where $\alpha = \alpha^*$ minimises

$$\int_{D^*} (h_{opt}(t) - h_{t,\alpha})^2 dt \tag{15}$$

with

$$h_{opt}(t) = (g''(t) - f''(t, \theta_t^*) + 2\beta_i(t))^{-\frac{2}{5}} k_2^{-\frac{2}{5}} n^{-\frac{1}{5}} K_1^{\frac{1}{5}} g(t)^{\frac{1}{5}}, \tag{16}$$

and $\beta_i(t)$, K_1, k_2 and θ_t^* as defined in section 4.2. As previously stated, our definition of the latter involved a large n approximation, but once again this method appears to work equally well with smaller sample sizes. Differentiating with respect to α, and setting the resulting equation equal to zero, we find that α^* solves

$$\int_{D^*} \log \Upsilon(t) h_{t,\alpha}(h_{opt}(t) - h_{t,\alpha}) dt = 0 \tag{17}$$

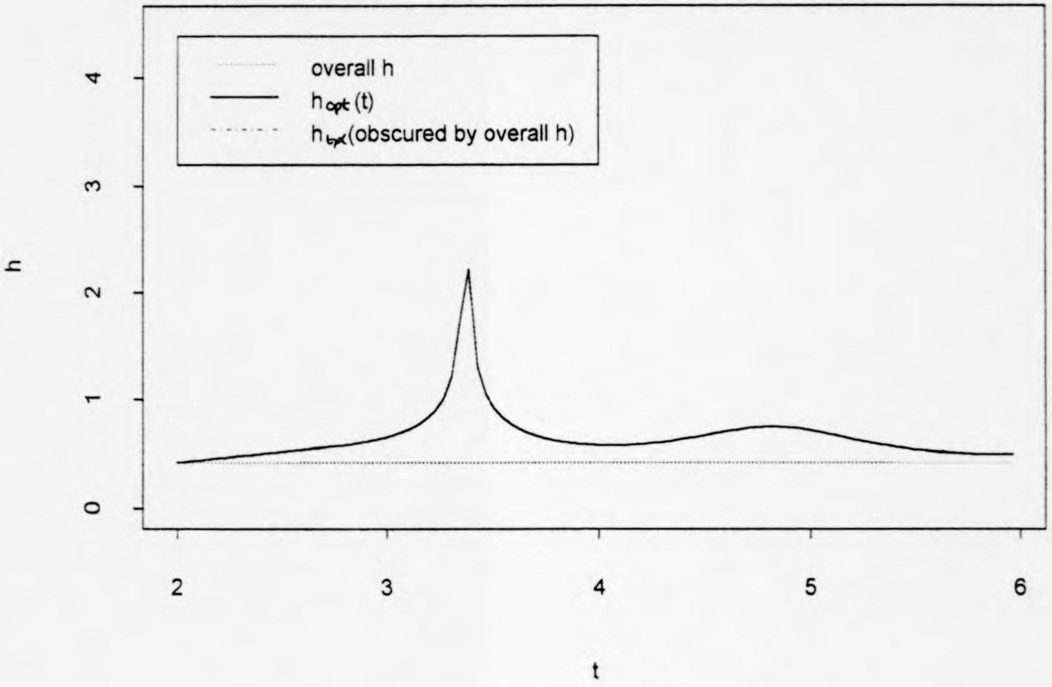
which, if h is chosen automatically, is equivalent to

$$\begin{aligned} &\int_{D^*} \log \Upsilon(t) \Big(\Upsilon(t)^{-\alpha} \left(g''(t) - f''(t, \theta_t^*) + 2\beta_i(t)\right)^{-\frac{2}{5}} g(t)^{\frac{1}{5}} \\ &- (g''(t^*) - f''(t, \theta_{t^*}^*) + 2\beta_i(t^*))^{-\frac{2}{5}} g(t^*)^{\frac{1}{5}} \Upsilon(t)^{-\alpha} \Big) dt = 0. \end{aligned} \tag{18}$$

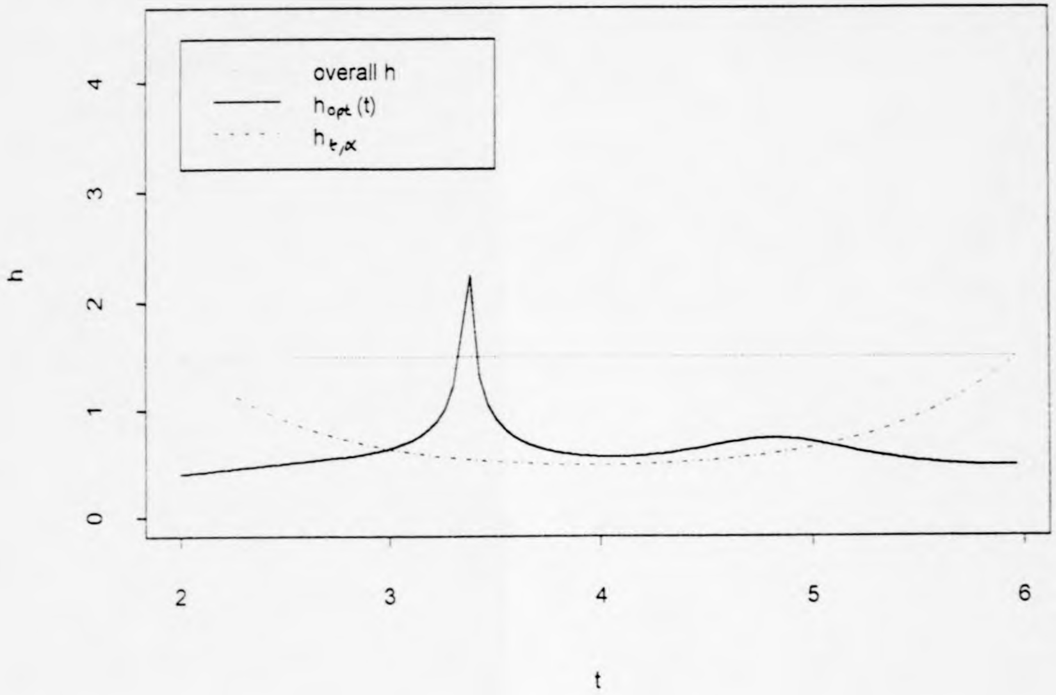
The fundamentals of this idea are well illustrated by plots 5a, 5b and 5c. In these three cases we are calculating $h_{t,\alpha}$ having chosen to fit a Normal distribution to a set of data, hence the two boundary points. Given our function $h_{opt}(t)$ over D^* and a fixed overall bandwidth h, then applying adaption pulls $h_{t,\alpha}$ away from h and towards $h_{opt}(t)$. When h is small, as it is likely to be if chosen by the automatic method, α^* will be very close to 0, or possibly less than 0 giving us a best choice of $\alpha = 0$ (see plot 5a). As h increases in order to minimise the distance between $h_{t,\alpha}$ and $h_{opt}(t)$, a lot of adaption is required to 'pull down' $h_{t,\alpha}$ from h (see plot 5b).

Yet again we find that this idea works well for small h and improves on previous methods in its treatment of medium sized h values. It retains the 'invariance' advantages of methods (i) and (iii) in that sample size and type of kernel do not affect our choice of α. As well as this it has an advantage over methods (i) and (iii) with respect to the thorny problem concerning the unknown second derivative of the true distribution. In method (i) (see equation (4)) and method (iii) (see equation (13)), our choice of α depends on $g(t)$ and $g''(t)$, both to the powers of magnitude 2. Preliminary estimates of these values are provided by kernel density estimation and are also required for the evaluation of θ_t^*. As explained in subsection 4.2.2, these estimates can

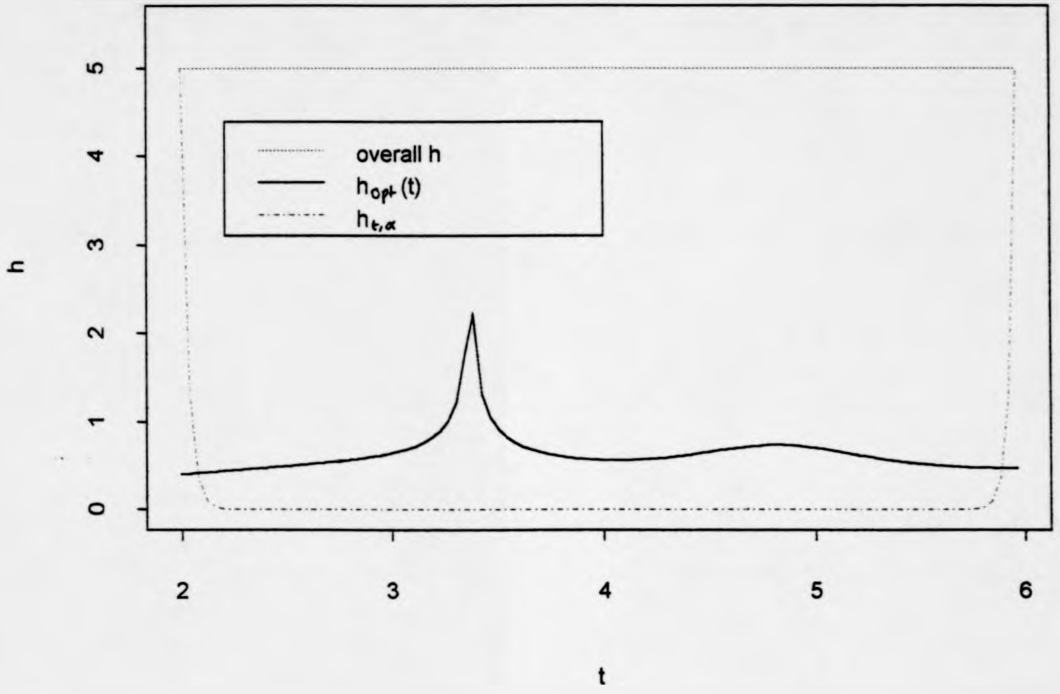
PLOT 5a: Choosing optimal α by minimising distance over D*
between $h_{opt}(t)$ and $h_{t,\alpha}$ (overall h small)



PLOT 5b: Choosing optimal α by minimising distance over D*
between $h_{opt}(t)$ and $h_{t,\alpha}$ (overall h medium)



PLOT 5c: Choosing optimal α by minimising distance over D^* between $h_{opt}(t)$ and $h_{t,\alpha}$ (overall h large, no restriction on $h_{t,\alpha}$)



be inaccurate when g is clearly non-Normal. While method (iv) still requires estimates of the true distribution and of its second derivative for use directly in equation (18), with the latter being liable to large random errors, the power to which they are taken has been reduced to $\frac{1}{5}$ and $-\frac{2}{5}$ respectively. Thus any inaccuracy in our kernel derivative estimate of $g''(t)$ will have less effect than it had in the previously considered methods. But like method (iii), method (iv) is not recommended for use with distributions such as the exponential, where the shape, discontinuities and bounded nature of the density function can lead to very inaccurate preliminary kernel estimates of $g(t)$, $g'(t)$ and $g''(t)$ in some regions of t. Most significantly, this will produce poor estimates of $h_{opt}(t)$ in these regions.

'Large h' problems also occur and are visually apparent in plot 5c. When h gets much bigger than the average of $h_{opt}(t)$, the value of α required to minimise equation (15) will increase dramatically, resulting in very small $h_{t,\alpha}$ values around the mode of f. This is more a structural problem than one caused by a poor approximation. Our value of $h_{t,\alpha}$ will always be equal to h at the boundary points, so as h increases a large amount of α is needed to pull $h_{t,\alpha}$ down to a value below h_{opt} elsewhere in D^*, in order to minimise (15). We can solve this problem by a placing a simple restriction on the size of α. Assume that difficulties caused by having too small a handpicked value of h are unlikely to occur, since it is improbable that we would want to deliberately cause noise in the tails by choosing a very small value of h. On the other hand, it is possible that we would want to use a much larger overall bandwidth than that chosen automatically, so as to ensure a smooth parametric estimate in the tails. Thus the problem of having too small a value of h should only arise from the automatic method. However, our automatically chosen overall bandwidth is the optimal local bandwidth value $h_{opt}(t)$ at the maximum difference boundary point $t = t^*$, so

$$h = h_{opt}(t^*) \geq \min_{D^*} h_{opt}(t).$$

Therefore we have a lower bound on our automatically chosen h, and are assuming that we would not want to handpick a smaller bandwidth than this. Overall bandwidth h is the least upper bound for local bandwidth $h_{t,\alpha}$.

Method (iv) involves extending this bound to $h_{t,\alpha}$. I suggest the following restriction on the best choice of α given in (18). Define the restricted best choice of α, method (iv), as

$$\min(\max(\alpha^*, 0), \bar{\alpha}). \tag{19}$$

Value $\bar{\alpha}$ of α is achieved when the smallest value of $h_{t,\alpha}$ equals the smallest value of $h_{opt}(t)$, $t \in D^*$, such that

$$\min_{D^*} h_{t,\alpha} = h \min_{D^*} \Upsilon(t)^{-\bar{\alpha}} = \min_{D^*} h_{opt}(t).$$

Equation (19) places a bound on how small $h_{t,\alpha}$ can become, by stopping α being chosen greater than $\bar{\alpha}$. We now translate this bound on $h_{t,\alpha}$ algebraically to the corresponding bound $\bar{\alpha}$ on α, writing the latter as

$$\bar{\alpha} = \frac{\log(h) - \log(\min_{D^*} h_{opt}(t))}{\min_{D^*} \Upsilon(t)}. \qquad (20)$$

For specific choices of parametric family f we can reduce this formula still further. Since λ is fixed for all t, $\min_{D^*} \Upsilon(t)$ will always occur at $\max_t f(t, \hat{\theta})$. If we know the mode of f, which will always be located in region D^*, then equation (20) simplifies even further. For example, when $f(t, \theta)$ is the density function of a Normal distribution, equation (20) reduces to

$$\bar{\alpha} = 2(\log(h) - \log(\min_{D^*} h_{opt}(t))),$$

and when $f(t, \theta)$ is the density function of an exponential distribution,

$$\bar{\alpha} = \log(h) - \log(\min_{D^*}(h_{opt}(t)).$$

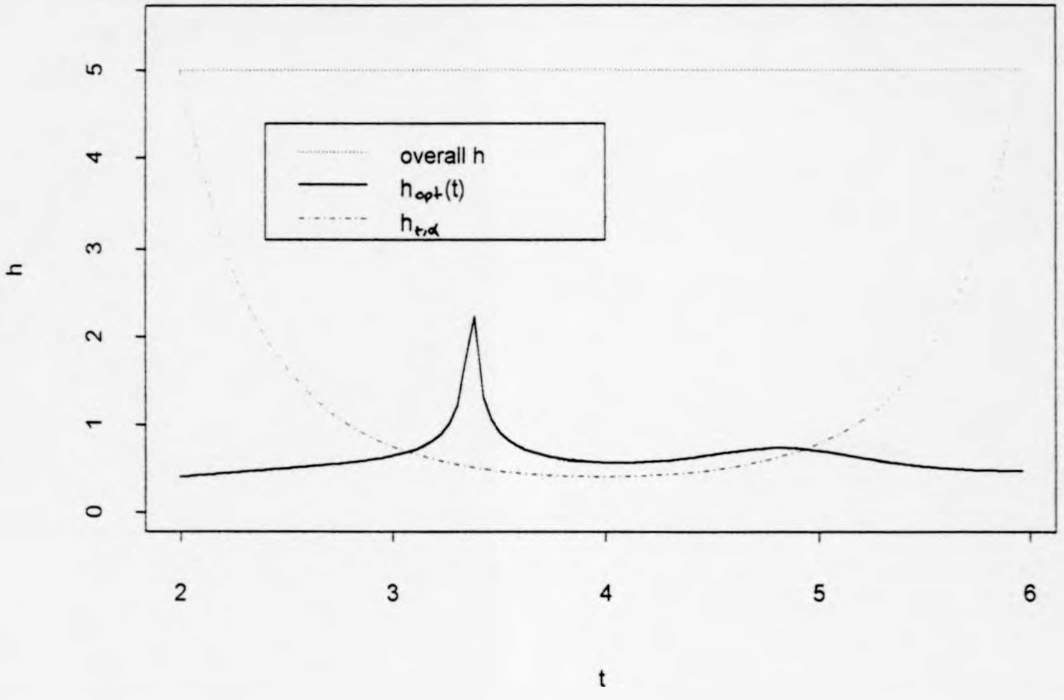
The value taken by $\min_{D^*} h_{opt}(t)$ can be found by using a simple minimisation program. In a practical case, where one is estimating the true distribution on a fine grid of points, it may be easier just to calculate $h_{opt}(t)$ for all these points in D^* and use the smallest value as an estimate of $\min_{D^*} h_{opt}(t)$.

Plot 5d illustrates method (iv) showing how $\min_{D^*} h_{opt}(t)$ creates an upper bound for α via $h_{t,\alpha}$, by preventing $h_{t,\alpha}$ from becoming very small around the mode of f when h is large. Compare this with what happens in plot 5c. Here no restriction is applied to the same example. We simply minimise (15), allowing a very large α to be chosen, which results in $h_{t,\alpha} \simeq 0$ for much of $t \in D^*$. The smallest value of $h_{t,\alpha}$ will always be at point t where $f(t, \hat{\theta})$ is largest which is the mode.

By placing a restriction on the minimisation of equation (15), we are attempting to ensure that $h_{t,\alpha}$ will never get so small that our adaptive semiparametric estimate in region D^* becomes just a series of spikes at the data points, as it does for very large h values in methods (i), (ii) and (iii). For a large range of examples considered, method (iv) worked very successfully. It required less computer time to select α than the other methods. For most examples, the restriction on the size of α came into action when overall bandwidth h was between 1 and 2 sample standard deviations.

For small to medium values of h it tended to choose slightly smaller values of α than the other methods. This too is an advantage, when we consider

PLOT 5d: Choosing optimal α by minimising distance over D* between $h_{opt}(t)$ and $h_{t,\alpha}$ (overall h large, restriction on $h_{t,\alpha}$)



the range over which α is chosen and the range over which it is applied. In methods (i) to (iii) we select α with the aim of minimising the MISE of the adaptive semi-parametric method over D^*. Since choosing α much greater than zero will take our adaptive semi-parametric estimate in D^{*C} towards our parametric estimate $f(t, \hat{\theta})$, in cases where f differs from the true distribution g we will incur a loss in the accuracy of our estimate in that region. As originally suggested in subsection 3.4.4, the best choice of α in terms of minimising the MISE over the whole range of t can be noticeably smaller than that chosen to minimise the MISE over D^*.

This problem is most likely to manifest itself in cases where the overall bandwidth h is of small to medium size. Our best choices of α over all t and over D^* will be virtually identical when h is very large, because our adaptive semi-parametric estimate will already approximate the parametric estimate in D^{*C} before any α is applied. When h is very small our best choice of α over D^* is likely to be zero. Refer back to section 3.4.4 and plot 3e for further details.

If h is so small that the estimate in D^{*C} is bumpy, it may be that the best choice of α in terms of minimising the MISE over all t is slightly greater than zero in order to smooth the tails. However, in this unlikely situation the best procedure would be to reselect a slightly larger value of h.

But as a general rule, our best choice of α for minimising the MISE over all t on which f is defined will always err on the small side of the best choice over D^*. Given that h and α are both large enough to avoid excessive noise in the tails, we will never expect it to be greater than the best choice of α over D^*. Any choice greater than this will by definition cause a loss of accuracy in D^*, and will also cause a loss of accuracy in D^{*C} for the reasons outlined above, **assuming that n is large and f is misspecified**. This is demonstrated in the examples given in section 5.3.

Unless we believe g to be extremely non-Normal, such as when it is an exponential distribution, then method (iv), which tends to give slightly smaller choices of α than methods(i), (ii) or (iii), is an effective method of automatically choosing a value of α for our adaptive semi-parametric method in terms of providing an accurate estimate of $g(t)$ for all t. It also has the necessary upper bound on α for when h is large.

5.3 Some examples of these methods in action

Noting that methods (i) to (iv) choose α with regard to a limited range of t, it is interesting to observe and compare these choices of α with each other, and with the 'observed best choice' when integrating over both the limited range D^* used in our methods and the more practically relevant full range of t. We define the **observed best choice** as that which minimises the observed ISE. It is possible to evaluate this when using a synthetic data set from a known true distribution g, with numerical integration used to calculate the ISE. A large sample size was used to reduce random error. I generated large synthetic data sets for different choices of distributions g and applied semi-parametric methods with different choices of f, a sensible yet incorrect parametric family chosen after observing a histogram of the data from true distribution g. The results from several of these are discussed in detail below. It is important to remember that using these synthetic data sets introduces sampling error, but they illustrate the above ideas, theory and effectiveness of the various methods quite well. Note that all plots are located after the three examples, on pages 167 to 181.

5.3.1 Example 1

I took a random sample of 5000 points from g, a bimodal distribution consisting of an equal mixture of two Normal distributions, which had means of plus and minus 2 respectively and identical variances equal to 1.5^2. Thus

$$g(t) = 0.5(g_1(t) + g_2(t)),$$

where

$$g_1 \sim N[-2, 1.5^2]$$

and

$$g_2 \sim N[2, 1.5^2].$$

Attempting to fit f, a Normal distribution, to this sample using the adaptive semi-parametric method, I produced plots 5e to 5g, showing the ISE's calculated over D^* and the full range of t for various values of α, for each of the 3 values of h. In other words I have plotted α vs 'ISE over D^*/'full range of t' for the adaptive semi-parametric method for 3 different choices of h, one the optimal choice via equations (16) of chapter 4, the others hand-picked. The 'automatic choices' of α were then calculated for each one of the methods outlined in this section, for each of the three h values (see table

5A). Observed best choices of α over both D^* and the whole range of t were visually apparent from the plots.

TABLE 5A

Automatic choices of α for example 1 under the four different selection methods (rounded to 2 decimal places)

h	Method (i)	Method (ii)	Method (iii)	Method (iv)
0.49	0.14	0.00	0.00	0.00
1.00	3.09	4.22	1.86	1.12
5.00	13.8	17.4	11.8	4.83

Plot 5e shows the observed best choice of α for small h. In this case the overall bandwidth of $h = 0.49$ was chosen by the automatic method of chapter 4. Sampling variability and slightly bumpy tails have caused the optimal α over all t to be just above zero, but all the automatic methods of selecting α work well. For the medium sized h value, $h = 1$, shown in plot 5f, the methods relying on approximations to the local kernel density estimate fare badly, choosing too large an h value. Method (iii) gets close to the best value over D^* but the best observed α over all t is nearest to that chosen by method (iv). For large h (see plot 5g), method (iv) is the only one to give a realistic choice of α, the 'restriction' idea working to good effect. All other methods choose α values so large that the density estimate produced consisted of a series of spikes at the data points in D^* and a purely parametric estimate in D^{*C}. Note that for large h, the best choice of α over D^* is almost identical to that over the full range of t, for the reasons suggested at the end of subsection 3.2.4.

Considering the observed ISE's in plots 5e to 5g, the combination of h and α which produces the smallest loss in accuracy occurs in plot 5e, where h is chosen by my automatic method of chapter 4, equation (16), and the observed best α is very close to any of those chosen by the various automatic methods of choosing α. Plot 5h shows the result of taking the automatic choice of $h = 0.49$, the corresponding best choice of $\alpha = 0$ under method (iv), and then producing an adaptive semi-parametric density estimate. Its accuracy to the true distribution compared to the performance of the parametric estimate is outstanding. Plot 5i compares our adaptive semi-parametric density estimate to a normalised histogram of the data and an ordinary kernel density estimate, demonstrating both how we avoid the oversmoothing of the latter method, and the loyalty to the data of the adaptive semi-parametric method. When constructing this kernel estimate, the bandwidth was chosen using the simple 'plug-in' formula from Silverman (1986), stated in equation (12) of chapter 4. Using a smaller bandwidth to increase accuracy around the modes lead to slight roughness in the tails of our kernel estimate.

5.3.2 Example 2

On this occasion the true distribution is Gamma[1,4], from which 500 random points were sampled. We again attempted to fit a Normal distribution to this. The procedure used in example 1 was then followed exactly. Plots 5j to 5l show results calculated over D^* and the equivalent results over the whole range of t. Table 5B gives the automatic choices of α from methods (i) to (iv).

TABLE 5B

Automatic choices of α for example 2 under the four different selection methods (rounded to 2 decimal places)

h	Method (i)	Method (ii)	Method (iii)	Method (iv)
0.51	0.16	0.00	0.00	0.00
1.00	1.84	3.10	1.50	0.72
5.00	12.1	14.2	10.1	4.61

The results are very similar to those from example 1. Our automatic methods all choose practical values of α when h is small ($h = 0.51$, chosen by the automatic method). Plot 5j illustrates that a choice of α around zero is appropriate if we want to minimise the ISE. For the larger values of h, only method (iv) gives choices of α close to the observed best choice over the full range. Our smallest observed ISE occurs when $h = 0.51$, and α is small. Plot 5m shows the resulting adaptive semi-parametric density estimate if we use $h = 0.51$ and α as chosen by method (iv). This is a visible improvement over the parametric estimate almost everywhere. In plot 5n this estimate is superimposed onto a normalised histogram to demonstrate its accuracy to the data. This is again excellent, though a slightly larger value of α may be preferable in order to smooth the small bump in the upper right tail. Plot 5j indicates that $\alpha = 0.2$ gives the smallest observed ISE. However this is a much smaller sample size than in example 1, so we should not expect an equivalent level of accuracy.

5.3.3 Example 3

For our third example we attempt to fit a Weibull distribution to data (sample size 1000) drawn randomly from true distribution g, where g is a mixture of two Gamma distributions. Its density function is defined as

$$g(t) = 0.5(g_1(t) + g_2(t)),$$

where

$$g_1 \sim \Gamma[1, 2.5]$$

and

$$g_2 \sim \Gamma[1, 5].$$

In this example, the distance between $f(t, \theta)$ and $g(t)$ is much smaller everywhere than it was in examples 1 and 2. In both left and right tails, the density function of the parametric family fitted to the data is almost coincident with that of the true distribution (see plot 5r). As a result of this, the observed ISE's of $f(t, \hat{\theta}_{t,\alpha})$ for the different choices of h and α are much smaller than in examples 1 and 2. The accuracy of the parametric fit to the true distribution in the tails also means that the majority of the observed ISE of $f(t, \hat{\theta}_{t,\alpha})$ over all t emanates from region D^{*}, where the density functions of f and g differ slightly in shape.

This is apparent in plots 5o, 5p and 5q. In plot 5o, h is chosen by the automatic method and is fairly small. Since $f(t, \hat{\theta})$ is already so close to $g(t)$ in the tails, as α increases there is little room for improvement there. The initial removal of noise in the tails of our estimate accounts for the observed best α over all t being slightly greater than 0; the best value evaluated over D^* only is equal to 0. In the region of high density D^*, increasing α above 1 causes a noticeable loss in accuracy due to our adaptive semi-parametric estimate becoming too non-parametric and spiky. There is a similar pattern for $h = 1$ and $h = 5$; the tail regions have little to no influence on the best observed choice of α as long as it is large enough to prevent any noise. For $h = 5$ (see plot 5q), the best observed choices of α for minimising the ISE over D^* and over all t are almost equal. The observed best choices of α over the two regions differ rather more for the 'medium-sized' $h = 1$ case (see plot 5p), but for both regions the observed ISE's stay at the same levels ($\simeq 0.00012$ and $\simeq 0.00025$ respectively) until α is greater than 2.5.

Method (iv) is still the best automatic method in terms of choosing α closest to its best observed value. All four methods again work well for the small value of h, but for the medium and large values, method (iv) is the only one which avoids choosing too large a value of α (see table 5C and compare with plots 5o, 5p and 5q).

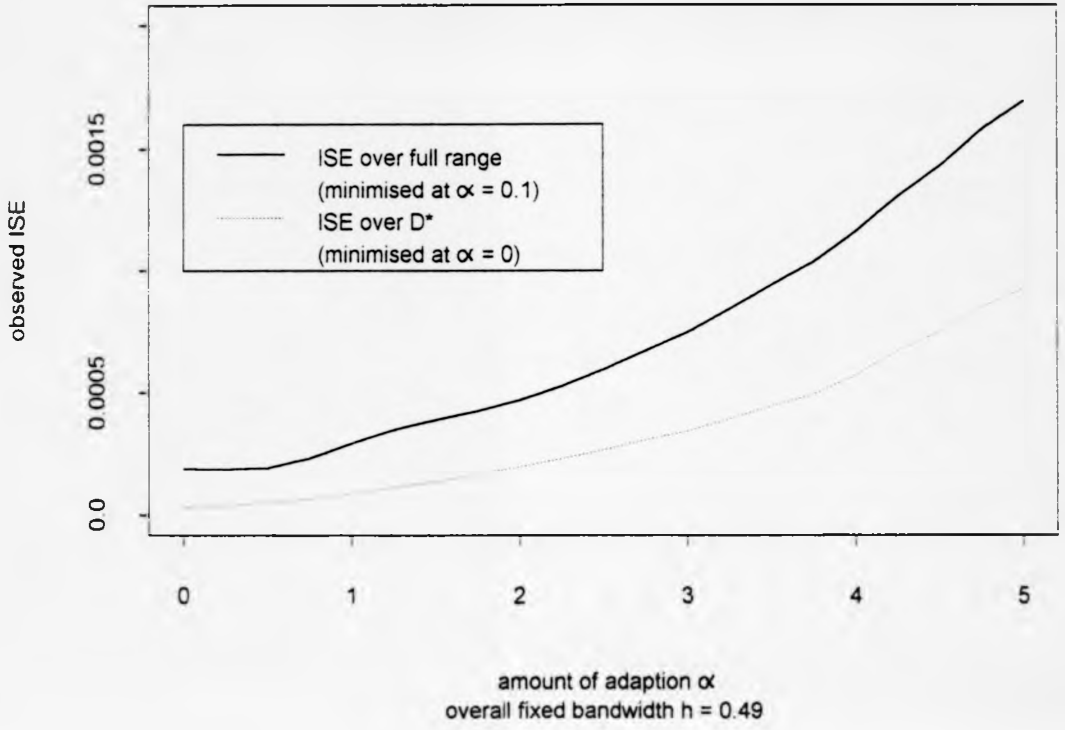
TABLE 5C

Automatic choices of α for example 3 under the four different selection methods (rounded to 2 decimal places)

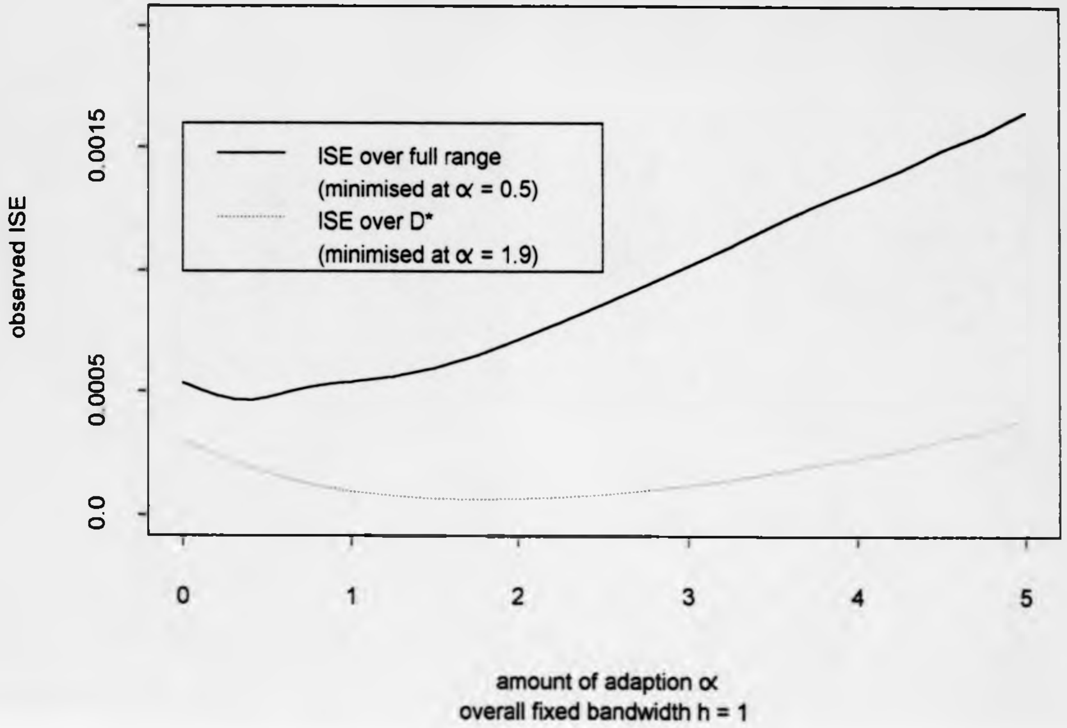
h	Method (i)	Method (ii)	Method (iii)	Method (iv)
0.55	0.58	0.00	0.00	0.00
1.00	3.11	6.10	0.90	0.27
5.00	13.6	8.99	13.6	5.33

Over the three choices of h and the range of α values used alongside them, the smallest observed ISE over all t, which equalled 0.00021, occurred when $h = 0.55$, α=0.25 and when $h = 1$, $\alpha = 0.28$. The combination of automatically chosen $h = 0.55$ followed by $\alpha = 0$, selected using method (iv), gives an observed ISE very close to this (it equalled 0.00022). Similarly, if we handpick the overall bandwidth $h = 1$, then method (iv) chooses α almost identical to the best observed value. Plot 5r shows the adaptive semi-parametric estimate resulting from the latter case, and plot 5s compares it to an ordinary kernel estimate (bandwidth chosen as for example 1) and a normalised histogram of the data. The adaptive semi-parametric estimate exhibits two advantages over the kernel density estimate. Firstly it does not take values greater than zero when t is less than zero, and secondly it does not oversmooth around the mode of the true distribution. Plot 5r shows the adaptive semi-parametric method's advantage in accuracy over the parametric estimate around the mode.

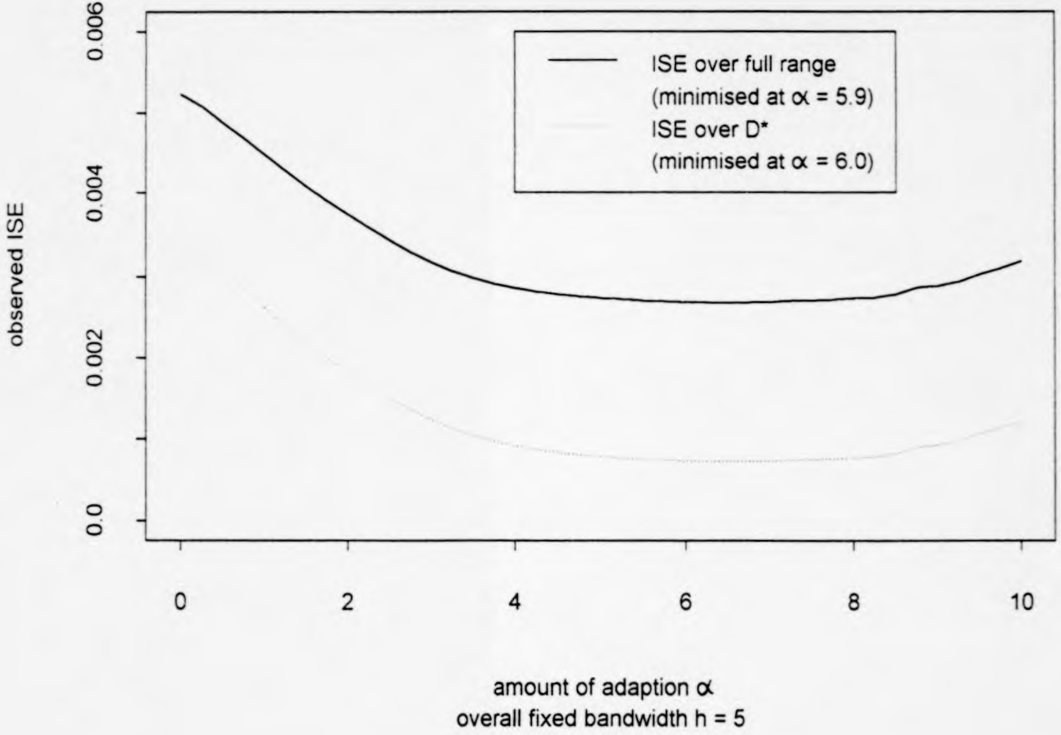
PLOT 5e: Observed ISE of semi-parametric method vs α, where g~bimodal and f~Normal



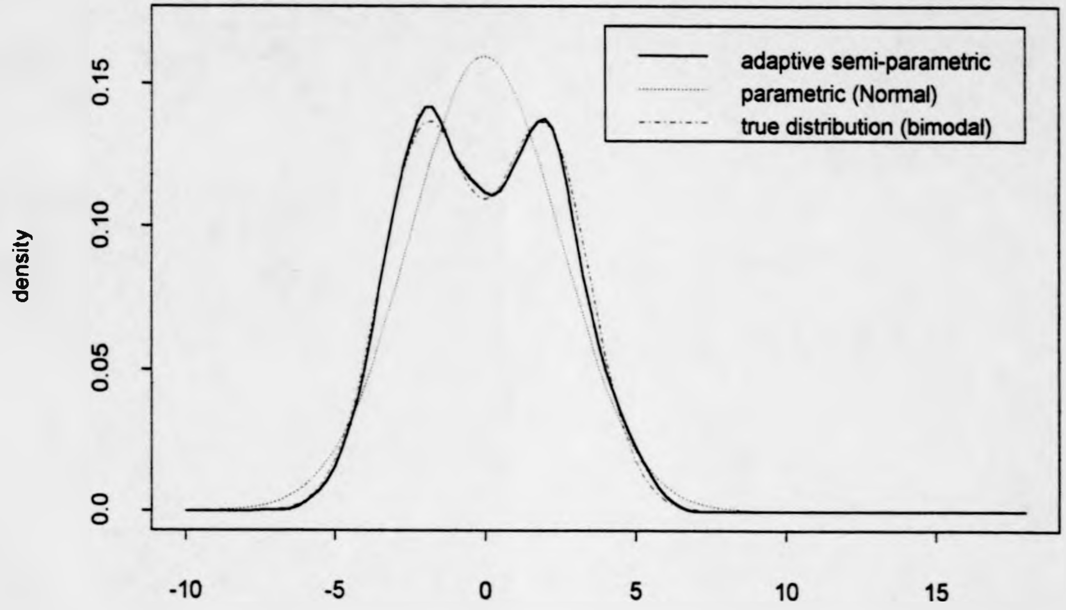
PLOT 5f: Observed ISE of semi-parametric method vs α, where g~bimodal and f~Normal



PLOT 5g: Observed ISE of semi-parametric method vs α, where g~bimodal and f~Normal

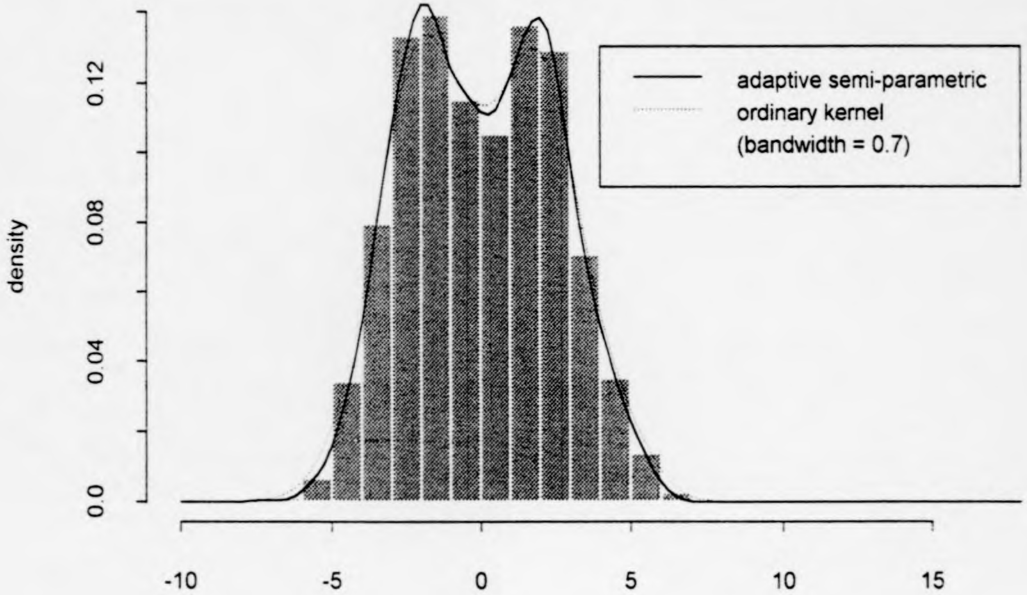


PLOT 5h: Comparing parametric and semi-parametric density estimates; semi-parametric uses optimal h and α



overall bandwidth h = 0.49   amount of adaption α = 0
random sample of 5000 points from bimodal distribution used

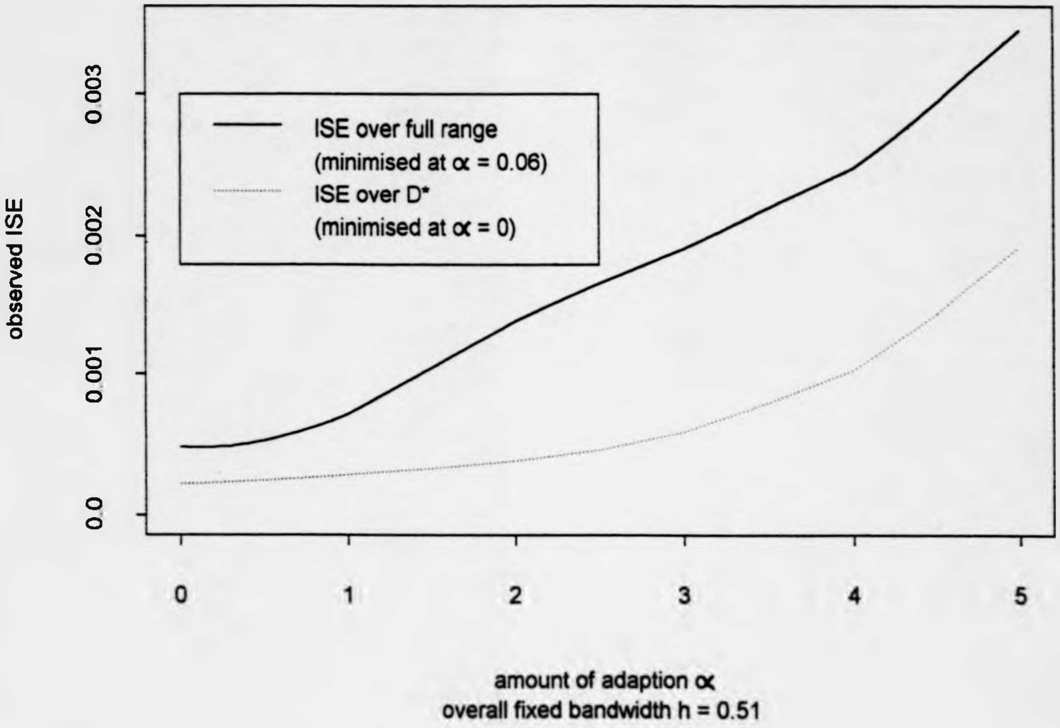
PLOT 5i: Comparing semi-parametric density estimate using optimal h and α with a normalised histogram of data and a kernel estimate



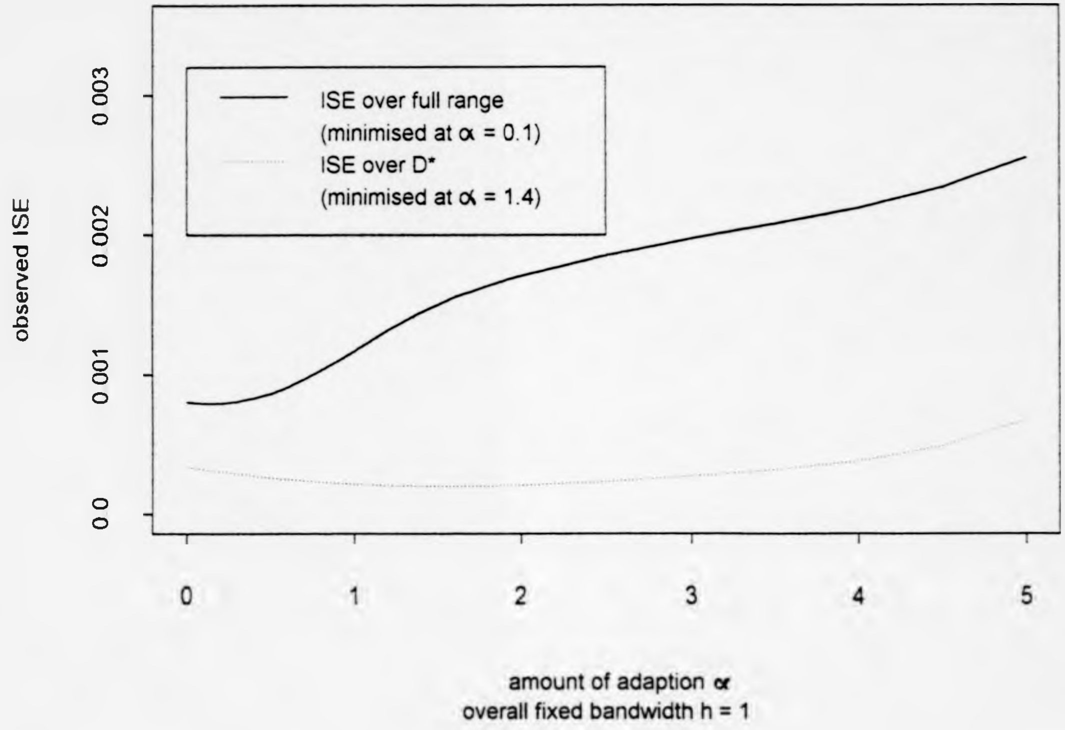
overall bandwidth h = 0.49 amount of adaption α = 0

random sample of 5000 points from bimodal mixture of Normals

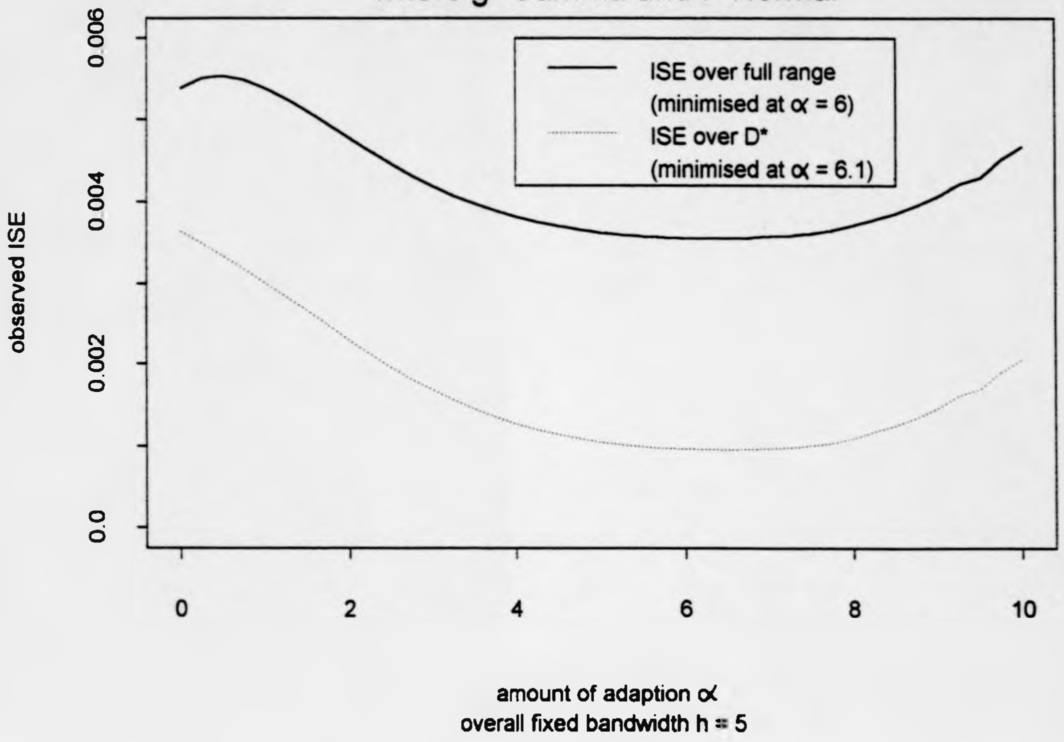
PLOT 5j: Observed ISE of semi-parametric method vs α, where g~Gamma and f~Normal



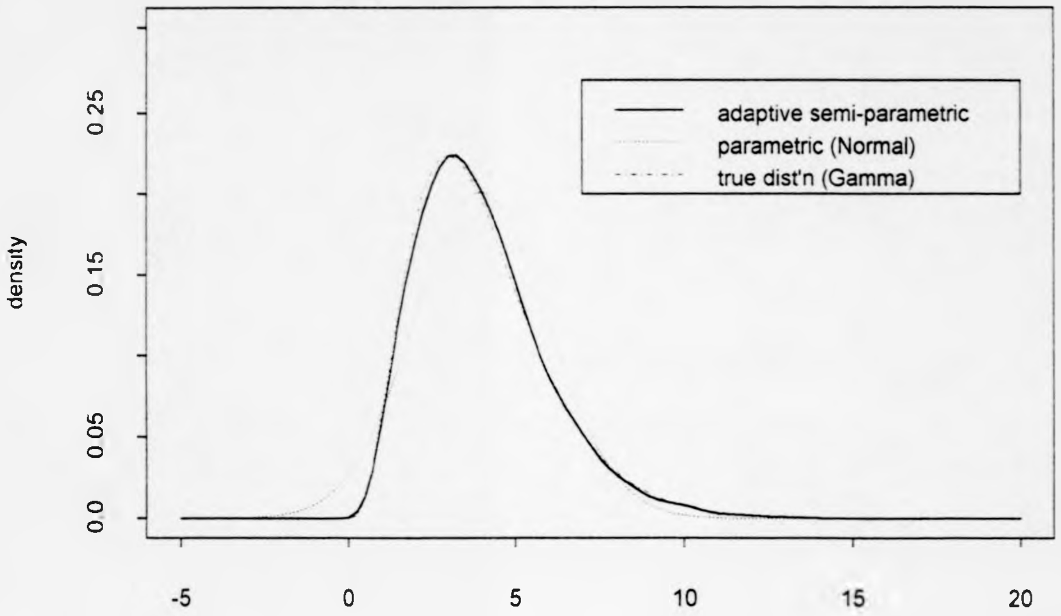
PLOT 5k: Observed ISE of semi-parametric method vs α, where g~Gamma and f~Normal



PLOT 5I: Observed ISE of semi-parametric method vs α, where g~Gamma and f~Normal

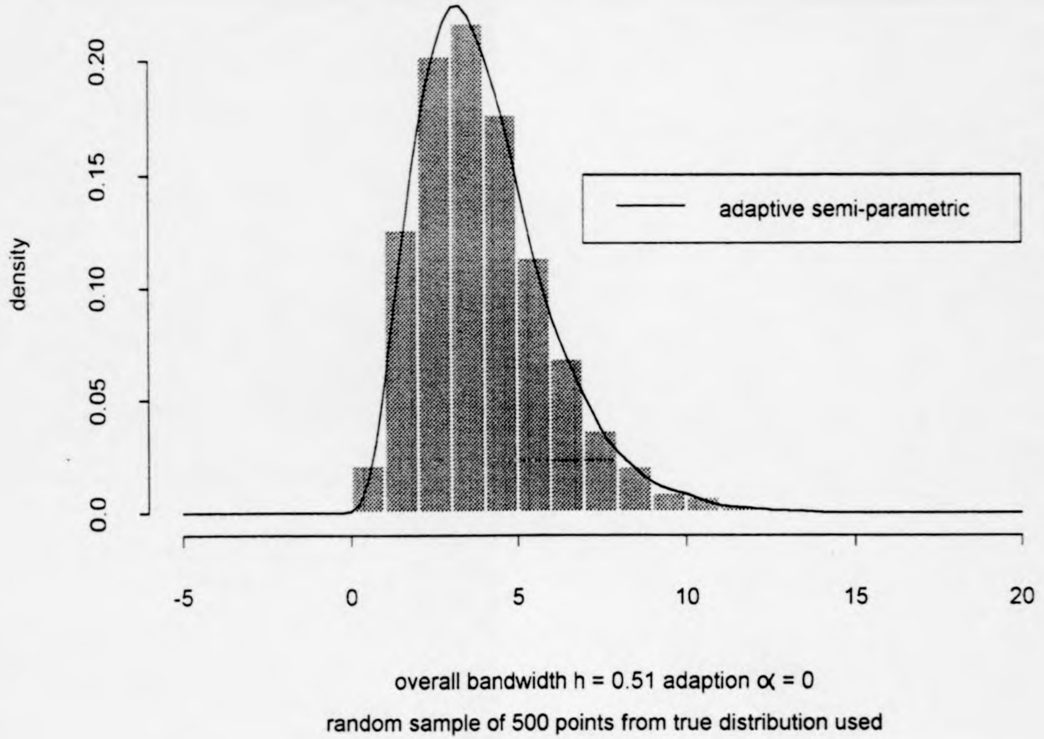


PLOT 5m: Comparing parametric and semi-parametric density estimates; semi-parametric uses optimal h and α

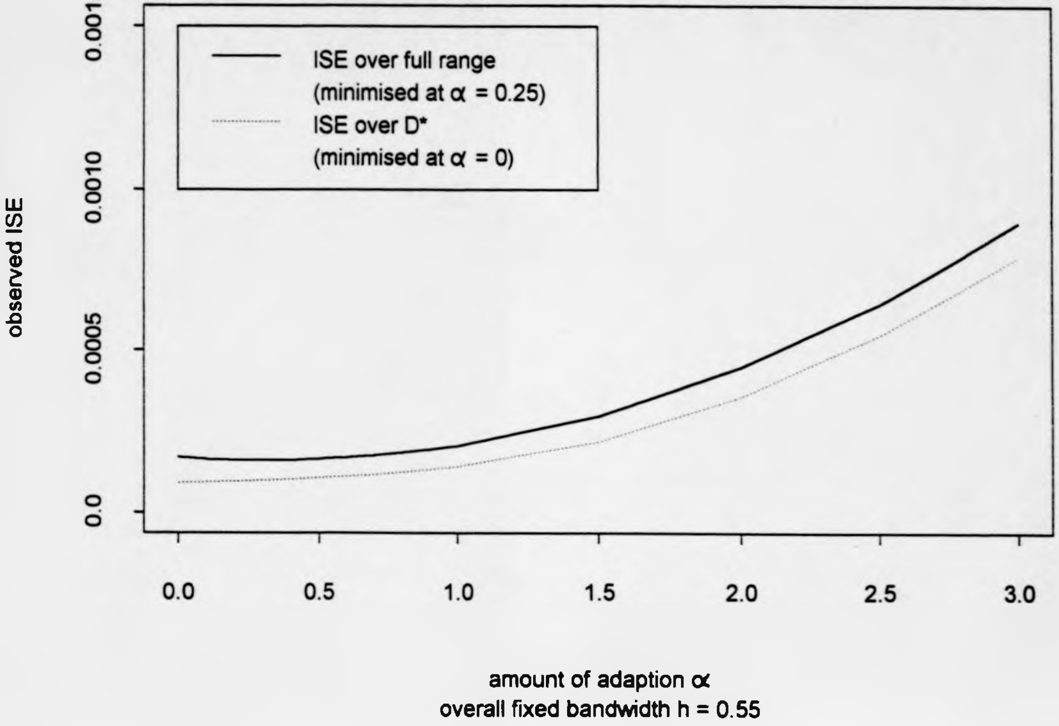


overall bandwidth h = 0.51 amount of adaption $\alpha = 0$
random sample of 500 points from Gamma(1,4) distribution used

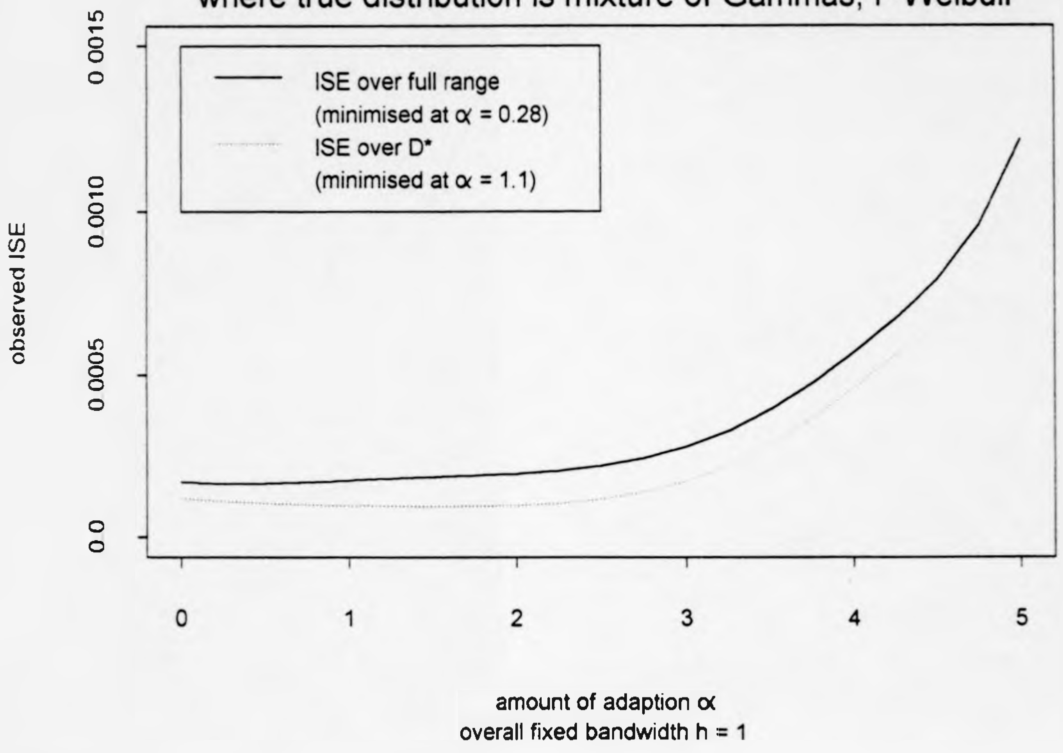
PLOT 5n: Comparing semi-parametric density estimate using optimal h and α with a normalised histogram of the data



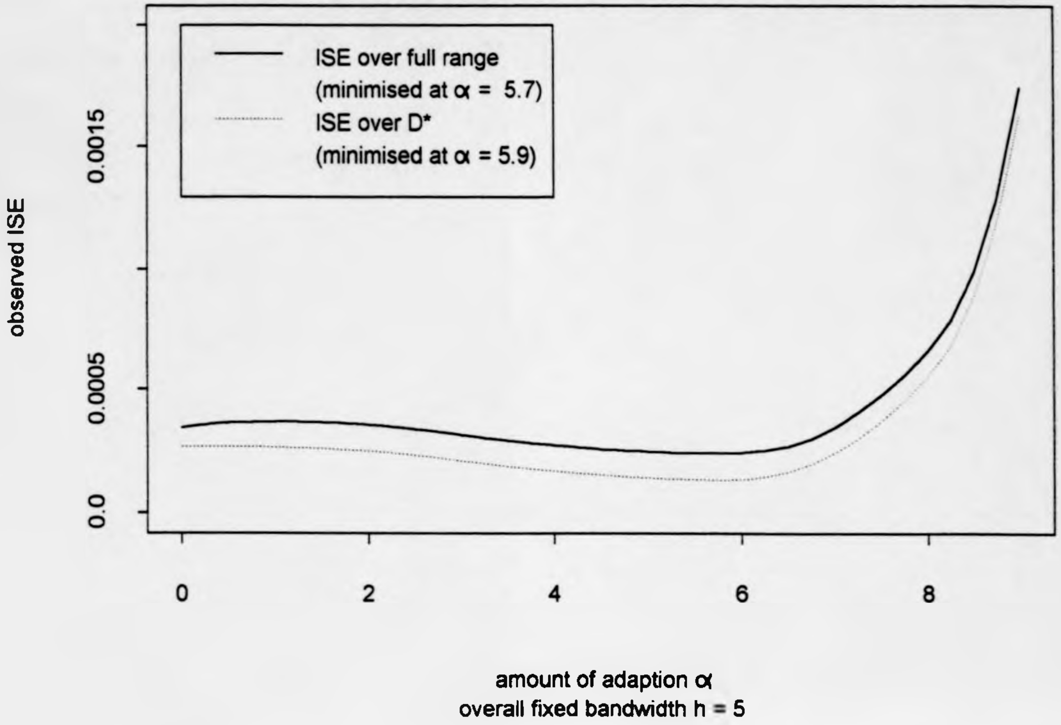
PLOT 5o: Observed ISE of semi-parametric method vs α, where true distribution is mixture of Gammas, f~Weibull



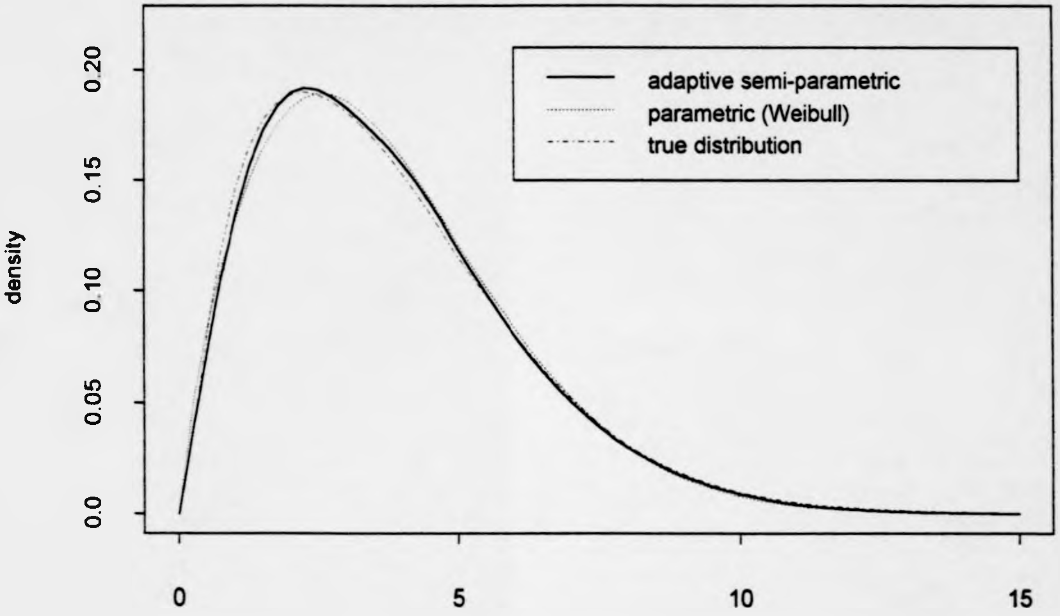
PLOT 5p: Observed ISE of semi-parametric method vs α, where true distribution is mixture of Gammas, f~Weibull



PLOT 5q: Observed ISE of semi-parametric method vs α, where true distribution is mixture of Gammas, f~Weibull

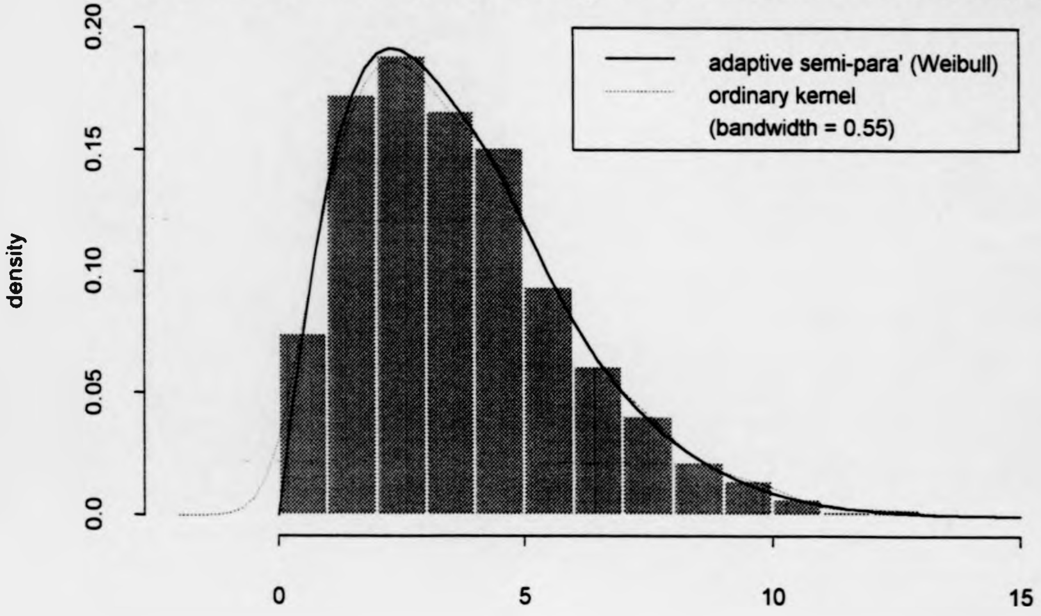


PLOT 5r: Comparing parametric and semi-parametric density estimates, semi-parametric uses optimal h and α



overall bandwidth h = 1   amount of adaption α = 0.28
random sample of 1000 points from mixture of Gamma distributions used

PLOT 5s: Comparing semi-parametric density estimate using optimal h and α with a normalised histogram of data and a kernel estimate



overall bandwidth h = 1 adaption α = 0.28

random sample of 1000 points from mixture of Gammas

5.4 Selecting h and α - a summary

Given that it was designed with parallel use of an adaptive constant α in mind, my selection procedure for h given in chapter 4 appears to work well on its own. In many of the practical examples attempted, I found that even before attempting to choose α and apply some adaption, I could obtain a good density estimate $f(t, \hat{\theta}_{t,\alpha})$ with h as chosen and $\alpha = 0$. It was very often the case that as in the simulated examples 1 to 3 of section 5.3, the automatic methods would choose a small h, and $\alpha \simeq 0$. Despite being dependent on the differences between f and g at only one point, and having been derived via large n and small h approximations this method of choosing h has proved effective over a wide range of examples. Any problems that have occurred have been limited to small samples from extremely non-Normal distributions such as the exponential, when the value of h chosen has occasionally been too small to smooth the tails or provoke the selection of any adaption when combined with the automatic methods of choosing α. Various methods for dealing with these cases are given in chapter 6. Depending on the situation, the automatic value can be used as given or, especially in small sample cases where our estimates of $g(t)$ and of its derivatives will be less reliable, as a rough guide to what a suitable size for h would be.

Of the four suggestions for choosing adaption parameter α, the first three, which all attempt to minimise small $h_{t,\alpha}$ approximations to the MISE of $f(t,\hat{\theta})$ over D^*, gave good choices of α until we began to contradict the small $h_{t,\alpha}$ criterion, when performance became poor. The alternative method (iv) chooses α to minimise the distance between $h_{t,\alpha}$ and the small h approximation to the optimum local bandwidth at t, over a restricted range of t. This enabled the introduction of an upper bound on the size of α, however large the overall bandwidth h. As h increased, only method (iv) gave sensible results. For cases where a histogram of the data suggests that the true density function is exceptionally non-Normal, bounded or discontinuous, method (ii) is recommended (as long as the overall bandwidth h is not too big), since it relies least on preliminary kernel derivative estimation. However it still relies on an initial small $h_{t,\alpha}$ approximation from the adaptive semi-parametric method to the local kernel estimate, and is therefore not foolproof!

It must be noted that the above ideas are only a limited range of suggestions; also that a computer is required to evaluate our 'best' h's and α's, though the minimisation routines used are simple and fast. A further advance would be to develop plug-in formulae for h and α, of equal simplicity to that suggested in Silverman (1986) for choosing the bandwidth for ordinary kernel density estimation.

6 Choosing h with reference to prior belief

6.1 Introduction

Like all automatic procedures, the methods for choosing h and α suggested above may not always be satisfactory. Method (iv), outlined in subsection 5.2.4, incorporates a safety net enabling us to avoid poor automatic choices of α. We now develop several safeguards for use when choosing h.

Since it is derived using small h and large n approximations of, rather than directly from the MSE of the semi-parametric method, our suggested automatic choice of h first given in equation (16) of chapter 4,

$$h = (g''(t^*) - f''(t^*, \theta^*_{t^*}) + 2\beta_i(t^*))^{-\frac{2}{5}} k_2^{-\frac{2}{5}} n^{-\frac{1}{5}} K_1^{\frac{1}{5}} g(t^*)^{\frac{1}{5}}, \qquad (1)$$

cannot be guaranteed to work well if the small h and large n criteria are violated. Equation (1) also requires estimation of $g(t)$ and its first two derivatives at boundary point $t = t^*$, which introduces a further margin for error. As a result of this, principally in small sample cases where we believe $g(t)$ to extremely non-Normal in shape, our automatic selection of h has occasionally been too small. This results in a noisy, largely non-parametric estimate before adaption is applied. Since h is already small in the areas of high density, our automatic methods of choosing the amount of adaption will select α close or equal to zero. Some noise in the tails will remain, unless we handpick a larger value of α which, while smoothing $f(t, \tilde{\theta}_{t,\alpha})$ when $t \in D^{*C}$, will result in very small values of $h_{t\alpha}$ for $t \in D^*$, making our estimate in this region even more volatile.

Alternatively, even when h is large enough to give reasonably smooth tails, we may believe that the true density is very close to that defining our parametric family f in the tails. Therefore we will want to use an even larger bandwidth and select a large value of α. Such situations motivate a procedure for increasing the size of overall bandwidth in line with our prior beliefs about the properties of the sample, without having to resort to handpicking h.

It must be emphasised that the methods given in this chapter are simply suggestions for dealing with problems I have occasionally encountered when applying the ideas of chapters 4 and 5. In the vast majority of examples tackled, choosing h as recommended in chapter 4 combined with α chosen by method (iv) from chapter 5 has proved satisfactory, at least as a rough indication of the optimal values of h and α.

6.2 A method for incorporating prior belief

Consider the formula for our best choice of overall bandwidth h, most recently given in equation (1). We will argue in section 6.3 that an effective way of incorporating prior belief leads to replacing the first and second derivatives of the true distribution at boundary point t^*, $g'(t^*)$ and $g''(t^*)$, by

$$(1-p)g'(t^*) + pf'(t^*, \theta^*_{t^*}) \tag{2}$$

and

$$(1-p)g''(t^*) + pf''(t^*, \theta^*_{t^*}) \tag{3}$$

respectively, where $0 \leq p < 1$.

This adjustment to equation (1) replaces

$$\beta_i(t^*) = \left(\frac{\frac{d}{d\theta_i} f'(t^*, \theta^*_{t^*})}{\frac{d}{d\theta_i} f(t^*, \theta^*_{t^*})} - \frac{f(t^*, \theta^*_{t^*})}{g(t^*)} \right) (g'(t^*) - f'(t^*, \theta^*_{t^*}))$$

by

$$\beta^*_i(t^*) = \left(\frac{\frac{d}{d\theta_i} f'(t^*, \theta^*_{t^*})}{\frac{d}{d\theta_i} f(t^*, \theta^*_{t^*})} - \frac{f(t^*, \theta^*_{t^*})}{g(t^*)} \right) (1-p) (g'(t^*) - f'(t^*, \theta^*_{t^*}))\,.$$

Note that $\beta^*_i(t) = 0$ when θ is a vector. Subsection 4.2.1 gives further details. These changes now give an optimal choice of h defined as

$$h' = (1-p)^{-\frac{2}{5}} (g''(t^*) - f''(t^*, \theta^*_{t^*}) + 2\beta^*_i(t^*))^{-\frac{2}{5}} k_2^{-\frac{2}{5}} n^{-\frac{1}{5}} K_1^{\frac{1}{5}} g(t^*)^{\frac{1}{5}}. \tag{4}$$

From now on we will consider this to be the formula for automatic selection of the overall bandwidth unless otherwise stated.

As p increases from 0, at which our formula for choosing h is unchanged from (1), towards its upper limit of 1, our choice of h will become larger. Given that $f \neq g$, we now require an increased value of α to achieve optimum accuracy to the true distribution in D^*. (As before we consider optimum accuracy to have been attained when the MISE over the region in question is minimised). These larger values of α and h will cause $f(t, \hat{\theta}_{t,\alpha})$ to move towards $f(t, \hat{\theta})$ when $t \in D^{*C}$.

Parameter p can be thought of as a smoothing index related to the likely proximity of $f(t, \theta)$ to $g(t)$. We could handpick p with regard to our personal prior belief, but it would be useful to have an automatic selection procedure, at least as a guide to an appropriate choice of p. Replacing $g'(t)$ and $g''(t)$ by formulae (2) and (3) is motivated by the theory of section 5.3, which shows how, after taking parametric family f as a prior distribution of g, the posterior expectation of our true density function given the data can be written in the form

$$(1-p)\hat{g}(t) + pf(t, \theta).$$

Equation (4) uses a similar arrangement. Though (2) and (3) are weighted sums of first and second derivatives of the density functions defining f and g, it seems reasonable to allow our choice of p to be related to the proximity of the density functions themselves. The value which p takes will affect our choice of bandwidth. This controls the 'placing' of our semi-parametric density estimate between the parametric and kernel density estimates.

6.3 The motivation for and formulation of p

We want parameter p to measure how well our parametric model f fits the data from true distribution g. To quantify the performance of f we consider g unknown with prior mean f, and choose p relative to how much the data suggests that g varies from this mean.

One way of approaching this is to suppose that, given data $x_1, ..., x_n$, then x_i, our ith observation, is equal to

$$\xi_i + \bar{h}\epsilon_i$$

where $i = 1, ..., n$, ξ_i are i.i.d $g(\xi_i)$ and the error ϵ_i are i.i.d. Normal[0,1]. Writing x_i in this way motivates the non-parametric kernel density estimate, if we think of each observation x_i being measured with standard deviation

$\bar{h}$. We assume that n is large and that the x_i's are measured with the error distributed Normally with mean 0 and variance $\bar{h}^2$. **Standard deviation $\bar{h}$ is small relative to the overall variance of the ξ_i's.** Then the posterior density function of ξ_i given x_i is approximately

$$\frac{1}{\bar{h}}\phi\left(\frac{x_i - \xi_i}{\bar{h}}\right).$$

Let n_t be the number of ξ_i's in the interval $(t, t + dt)$. We find that

$$\begin{aligned} E(n_t|X) &= E\sum_{i=1}^{n}(1 \text{ if } t < \xi_i < t + dt \text{ , 0 otherwise }) \\ &= \sum_{i=1}^{n} P(t < \xi_i < t + dt) = \frac{1}{h}\sum_{i=1}^{n}\phi\left(\frac{x_i - t}{h}\right) dt. \end{aligned}$$

Now take

$$g_t = P(t < \xi < t + dt) = g(t)dt,$$

and let the distribution of the vector of g_t's over a fine grid of width dt be Dirichlet(κ, f_t) where $f_t = f(t, \theta)dt$ and $f(t, \theta)$ is our chosen parametric density. Then the prior mean of g_t is f_t and the prior variance of g_t is $f_t(1 - f_t)(\kappa + 1)^{-1}$. Therefore, when κ is small, g_t varies substantially from f_t, but if κ is large, the variance is small. By using the resulting beta-binomial structure we can define our smoothing index in terms of κ, so that it is related to closeness of $f(t, \theta)$ to $g(t)$.

Since n_t given g_t is distributed binomially such that

$$n_t|g_t \sim \text{binomial}(n, g_t),$$

and

$$g_t \sim \text{beta}(\kappa f_t, \kappa(1-f_t)), \tag{5}$$

then this implies that

$$g_t|n_t \sim \text{beta}(\kappa f_t + n_t, \kappa(1-f_t) + n - n_t), \tag{6}$$

so that

$$E_g(g_t|n_t) = \frac{\kappa f_t + n_t}{\kappa + n}.$$

Hence

$$\begin{aligned} E(g_t|X) = E_{n_t}\left(E_g(g_t|X, n_t)\right) &= E_{n_t}.\left(\frac{\kappa f_t + n_t}{\kappa + n}|X\right) \\ &= \frac{\kappa f_t + \frac{1}{h}\sum_{i=1}^{n}\phi\left(\frac{x_i - t}{h}\right)dt}{\kappa + n}. \end{aligned}$$

But as $\frac{g_t}{dt} \sim g(t)$, $\frac{f_t}{dt} \sim f(t, \theta)$ and

$$\hat{g}(t) = \frac{1}{n\bar{h}}\sum_{i=1}^{n}\phi\left(\frac{x_i - t}{\bar{h}}\right),$$

which is the kernel density estimate of $g(t)$ using a Gaussian kernel and with bandwidth $\bar{h}$, then the posterior expectation of the true distribution is of the form

$$E(g(t)|X) = \frac{\frac{\kappa f(t,\theta)}{n} + \hat{g}(t)}{\frac{\kappa}{n} + 1}. \tag{7}$$

This is simply a weighted sum of the density function of f and a kernel estimate of the density function of g. Therefore we can write

$$E(g(t)|X) = (1-p)\hat{g}(t) + pf(t,\theta),$$

where

$$p = \frac{\kappa}{n+\kappa}. \tag{8}$$

Our choice of p is dependent on κ, and thus on the variance of g_t from its prior mean f_t. If we choose p as formulated in equation (8) for use in equation (4), it has the desirable properties of being small when the variance is large, thus retaining a small value of h, and it increases as the variance decreases. When we believe $g(t)$ to be very close to $f(t,\theta)$, then κ is large, giving a value of p close to 1. This ensures a substantially larger choice of h from equation (4) than that given by equation (1).

We have introduced κ as a parameter of the prior distribution of g, but in section 6.4 we consider an empirical Bayes argument by which κ can be estimated from the data itself. Section 6.5 suggests several methods for adapting this argument by looking ahead and prioritising different regions of the adaptive semi-parametric estimate which we are attempting to produce. These methods rely on the dependence of p on κ, h on p, and if the adaption parameter is chosen automatically, the dependence of α on h. The relationships can be summarised as follows;

$$\kappa \text{ is small (vague prior)} \;\Rightarrow p \text{ is small} \;\Rightarrow h \text{ stays small,}$$

or alternatively

$$\kappa \text{ is large} \;\Rightarrow p \text{ is large} \;\Rightarrow h \text{ increases .}$$

6.4 An estimate of κ

A possible procedure for estimating κ is to use the following empirical Bayes argument. We discretise the data into l class intervals $(\Delta_1, \Delta_2), \ldots, (\Delta_l, \Delta_{l+1})$, the choice of which shall be discussed later, and then define m_j to be the number of observations in the jth class interval (Δ_j, Δ_{j+1}) and

$$\psi_j = \int_{\Delta_t}^{\Delta_{j+1}} g(\xi)d\xi.$$

Since we are considering distribution f as a prior for true distribution g, from equations (5) and (6) we deduce that

$$\psi_j \sim \text{beta}\,(\kappa r_j, \kappa\,(1 - r_j))$$

where

$$r_j = \int_{\Delta_j}^{\Delta_{j+1}} f(t,\theta)dt.$$

Assuming that $\bar{h}$ is small relative to the size of class intervals $|\Delta_{j+1} - \Delta_j|$, from section 6.3 we have

$$m_j|\psi_j \sim \text{binomial}(n, \psi_j).$$

This implies that

$$E(m_j) = E\left(E(m_j|\psi_j)\right) = nE(\psi_j) = nr_j, \tag{9}$$

where r_j is the probability of an observation being in interval j given that the true distribution of the data is defined by our parametric family f. We can estimate this value by

$$\hat{r}_j = \int_{\Delta_j}^{\Delta_{j+1}} f(t,\hat{\theta})dt.$$

Now

$$\begin{aligned} Var\left(m_j\right) &= E\left(Var(m_j|\psi_j)\right) + Var\left(E(m_j|\psi_j)\right) \\ &= E\left(n\psi_j(1-\psi_j)\right) + Var\left(n\psi_j\right) \\ &= nr_j - nr_j{}^2 - n\frac{r_j(1-r_j)}{\kappa+1} + \frac{n^2 r_j(1-r_j)}{\kappa+1} \\ &= \left(1 + \frac{n-1}{\kappa+1}\right)\left(nr_j(1-r_j)\right) \end{aligned} \tag{10}$$

and so

$$\frac{Var\,(m_j)}{nr_j(1-r_j)} = E\left(\frac{(m_j-nr_j)^2}{nr_j(1-r_j)}\right) = 1+\frac{n-1}{\kappa+1}. \qquad (11)$$

The sum of the left hand side of this equation is similar to the expectation of $\frac{\chi^2}{l}$, defined as

$$\frac{\chi^2}{l} = \frac{1}{l}\sum_{j=1}^{l}\frac{(m_j-nr_j)^2}{nr_j},$$

in that it is a weighted sum of the squared differences between the observed frequency and expected frequency of observations, assuming that the true distribution of the data is given by parametric family f. We can estimate this expectation by approximating r_j by $\hat{r}_j$ and taking the sample mean over the l intervals, which we define as

$$Z^* = \frac{1}{l}\sum_{j=1}^{l}\frac{(m_j-n\hat{r}_j)^2}{n\hat{r}_j(1-\hat{r}_j)}. \qquad (12)$$

When choosing the size and number of class intervals, follow similar guidelines to those for the discretisation of data for calculation of the χ^2 statistic,

$$\chi^2 = \sum_{j=1}^{l}\frac{(m_j-nr_j)^2}{nr_j}.$$

This implies choosing as many class intervals as is possible under the restriction that nr_j is greater than about 5, ensuring that the denominator of the fraction in equation (12) never becomes too small. If n is large, then our intervals will be chosen small, r_j will be close to zero for all j, and lZ^* can be approximated by the χ^2 statistic. There is still a great deal of scope for the choice of these intervals, which is a disadvantage of this discrete procedure. Different choices of intervals $(\Delta_1, \Delta_2), ..., (\Delta_l, \Delta_{l+1})$ will produce different estimates of κ and therefore different choices of p.

After rewriting equation (9) in terms of κ, we have

$$\kappa = \frac{n - \left(\frac{var(m_j)}{nr_j(1-r_j)}\right)}{\left(\frac{var(m_j)}{nr_j(1-r_j)}\right) - 1},$$

which we estimate by

$$\hat{\kappa} = \frac{n - Z^*}{Z^* - 1}. \qquad (13)$$

An automatic value of p is now chosen by inserting formula (10) into equation (8). Our estimate of κ and hence our choice of p has the desired property of being directly related to how much the observed data differs from what we would expect to observe if their true distribution was our chosen parametric family f.

If f is a very bad fit to the data from true distribution g, then Z^* will be large, giving small values of $\hat{\kappa}$ and p. Our automatic methods of choosing h and α from chapters 4 and 5 both tend to give small values unless f and g are very close. In this case we will end up with the appropriate choice of a small bandwidth and little if any adaption. In cases where h is still very small, this confirmation of our original selection of h could be seen as a suggestion that f is a very bad fit to the data from g, and that we should try fitting a different parametric family.

Alternatively if the data indicates that $f(t,\theta)$ and $g(t)$ are very close almost everywhere, then the resulting small value of Z^* will lead to κ being large, p being near to 1, and we will choose a larger overall bandwidth h. This in turn will provoke a larger automatically chosen value of α, so that $f(t,\hat{\theta}_{t,\alpha})$ will approximate the parametric density estimate $f(t,\hat{\theta})$ in the tails. This is a suitable outcome, since the data suggests that $f(t,\theta)$ appears to be close to $g(t)$.

Assume that we consider our automatically selected bandwidth using equation (1) somewhat small. Then finding p as suggested and using equation (4) to give a new automatically chosen bandwidth, the dependencies of section 6.3 can be extended to

$$f \text{ bad fit to data from } g \Rightarrow Z^* \text{ large } \Rightarrow \hat{\kappa} \text{ small (vague prior) } \Rightarrow p \text{ small}$$

$$\Rightarrow h \text{ stays small } \Rightarrow \alpha \text{ stays small } \Rightarrow \text{ estimate largely non-parametric },$$

or

$$f \text{ good fit to data from } g \Rightarrow Z^* \text{ is close to } 1 \Rightarrow \hat{\kappa} \text{ large } \Rightarrow p \text{ is close to } 1$$

$$\Rightarrow h \text{ increases } \Rightarrow \alpha \text{ increases } \Rightarrow \text{ estimate becomes more parametric .}$$

One flaw with this method is that equation (13) will take values less than zero when Z^* is less than 1. Plot 6a of κ against Z^* for a sample size of $n = 1000$ also features the discontinuity which occurs when $Z^* = 1$. Equation (8) will give values of p outside $[0, 1)$ if our estimate of κ is less than zero, a scenario that we must therefore avoid. A simple solution is to replace κ in equation (8) by

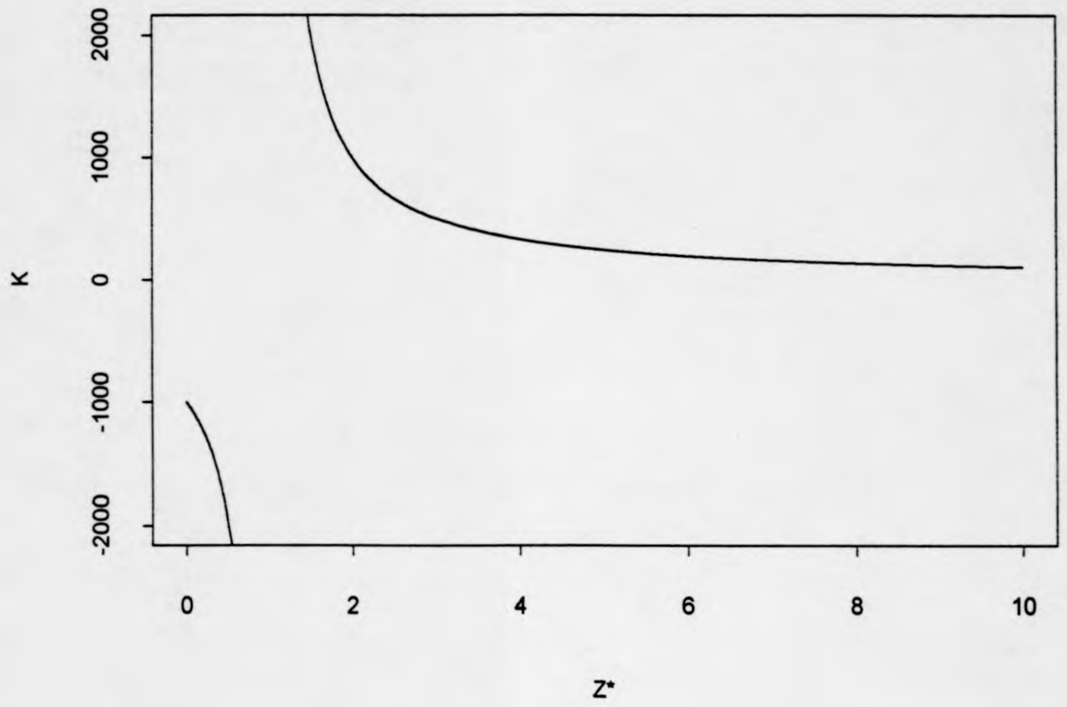
$$|\hat{\kappa}| = \left| \frac{n - Z^*}{Z^* - 1} \right| .$$

Plot 6b of the modulus of $\hat{\kappa}$ against Z^* shows the effect of this change.

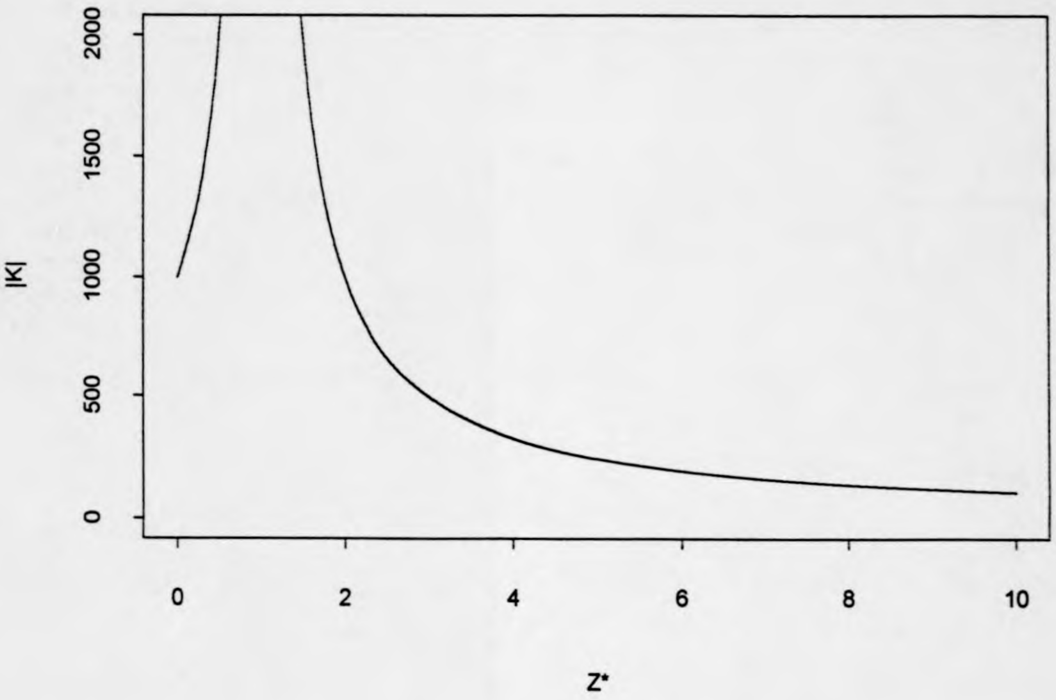
The problem of Z^* taking values in the interval $[0, 1]$, while awkward to solve, is not one that I have regularly encountered. Assuming that g differs from f enough to have motivated a small choice of h in the first place, it is unlikely that the data from g will be distributed closely enough to the expected frequencies under f to give a very small value of Z^*. In fact, given a fixed value of κ, if $f \neq g$, then as $n \rightarrow \infty$,

$$P(Z^* > 1) \rightarrow 1.$$

PLOT 6a: K vs Z* (n = 1000)



PLOT 6b: |K| vs Z* (n = 1000)



On the rare occasions that the value of Z^* is between 0 and 1, the idea of estimating κ by the modulus of $\hat{\kappa}$ is acceptable. It does not possess the intuitively desirable property of decreasing as Z^* increases; plot 6b shows how this breaks down when $0 \leq Z^* \leq 1$. However it does preserve the essential characteristic of being large whenever Z^* is small, giving a large value of p. This property is simple to show, since

$$Z^* \leq 1 \Rightarrow |\hat{\kappa}| \geq n,$$

therefore p estimated $|\hat{\kappa}|(|\hat{\kappa}| + n)^{-1}$ must always be greater than $\frac{1}{2}$ whenever Z^* is less than 1.

Another possible idea is to redefine p such that it takes a fixed value close to 1 whenever Z^* takes values in [0,1]. This lacks the simplicity of using $|\hat{\kappa}|$ and requires the subjective choice of a suitable value: should it be $p = 0.9$, $p = 0.99$ or $p = 0.999$?!

Problems will still occur when $Z^* = 1$! In any situation where $Z^* \in [0, 1]$, either handpicking a very large overall bandwidth h or even ignoring semi-parametric methods altogether in favour of parametric density estimation are equally logical and far simpler proposals!

6.4.1 An example

Selecting values of κ, p and thus h as suggested in above worked fairly well for a number of examples. The case given below is a typical situation in which introducing p is useful.

We return to the deer line transect data previously encountered and explained in fuller detail in sections 1.5 and 3.3. The data consists of perpendicular distances of deer sightings from a straight line. A histogram of the data, given in plot 1b, suggested that an exponential fit is plausible, but some local influence will be necessary around the mode. Constructing a parametric density estimate confirmed that the exponential distribution fits the tails very well indeed. When we used adaptive semi-parametric density estimation on this data set in chapter 3, we handpicked values of h and α, but now we have the tools necessary to attempt an optimal selection of these parameters.

First we try the methods given in chapters 4 and 5. Selecting h automatically without prior belief (using equation (1) from this chapter), we get a small overall bandwidth of $h = 1.01$. We now use methods (i) and (ii) of choosing α from subsections 5.2.1 and 5.2.2 respectively (since methods

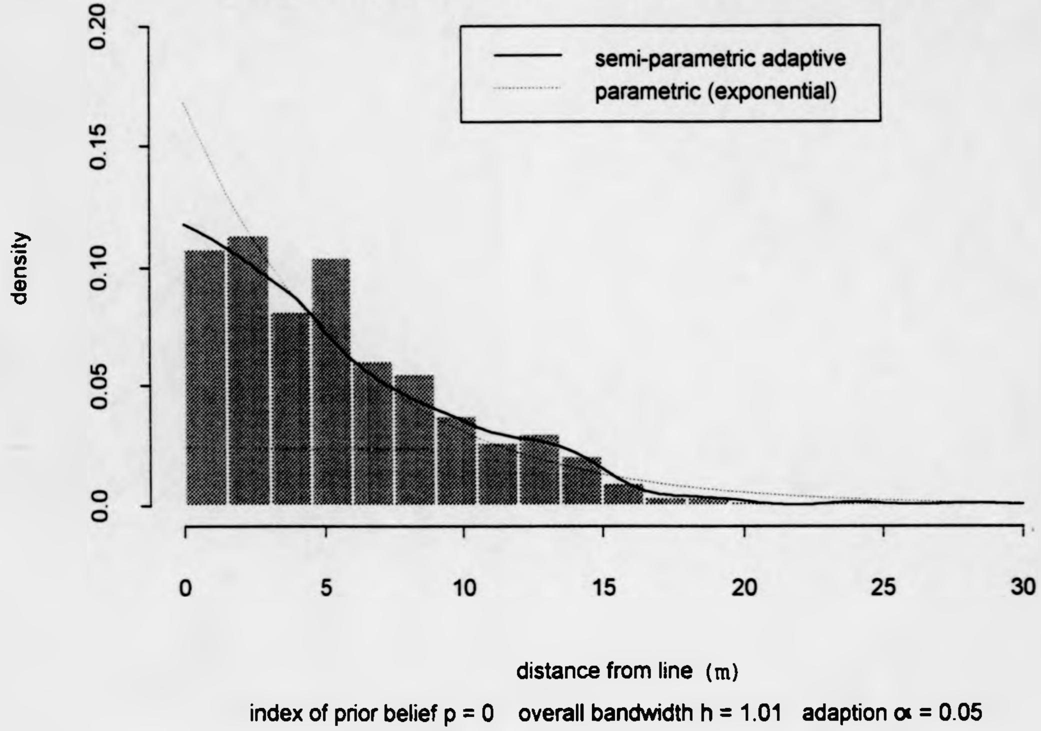
(iii) and (iv) are particularly unsuitable for use with exponential distributions due to problems in accurately estimating derivatives and finding $h_{opt}(t)$ around $t = 0$). Both methods used gave very similar small values of α; we will take $\alpha = 0.05$ as chosen by method (i). An adaptive semi-parametric estimate using these choices of h and α is shown in plot 6c. Our density estimate is adequate in high density regions where there is a lot of data, but its non-parametric nature results in poor performance in the right tail, where we get bumps at the data points.

Changing our method of selecting h to that of equation (4), thus incorporating prior belief, after calculating κ and p we find that h is increased to a value of 1.6. This in turn provokes a larger value of α. The resulting density estimate is shown in plot 6d. The right tail is now smoothed adequately, and our density estimate near the mode is neither too noisy nor too large.

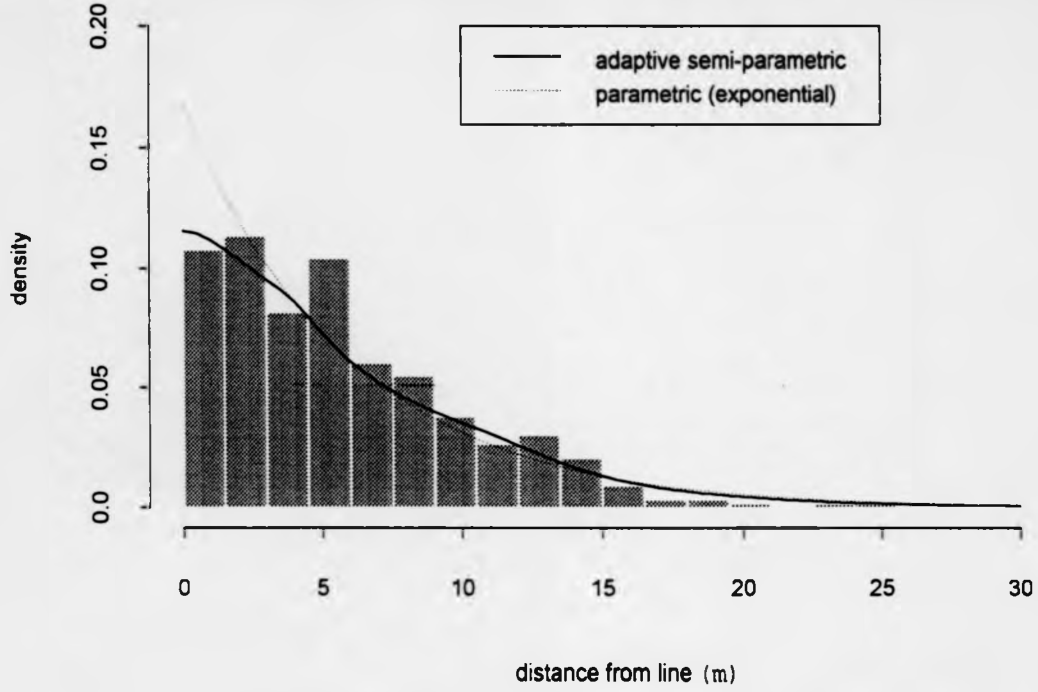
6.5 Two possible improvements for special cases

In section 6.4 the value of Z^* depends on equally weighted differences between observed and expected frequencies under parametric model f, over the range of the l intervals. We can adapt the method of sections 5.2 to 5.4 by weighting

PLOT 6c: Adaptive semi-parametric density estimate of deer line transect data with no `prior belief' incorporated



PLOT 6d: Adaptive semi-parametric density estimate of deer line transect data using index of `prior belief'



index of prior belief p = 0.7 overall bandwidth h = 1.6 adaption α = 0.72

the Z^* function so that differences between $f(t,\theta)$ and $g(t)$ in certain regions carry more weight than they do in others.

6.5.1 A tail-weighted method of estimating κ

For example, the initial motivation for introducing p was to avoid the bumpy tailed estimates caused by the overall bandwidth being too small. When selecting h, we are least concerned about our final estimate of the true distribution in areas of high density. Our adaptive semi-parametric estimate should be good here provided h is not very small, since we choose α to maximise the accuracy of $f(t,\hat{\theta}_{t,\alpha})$ to $g(t)$ over the range of $t \in D^*$. However, when h is small initially, this restriction to $t \in D^*$ means that our automatic methods will choose small values of α and the tails of our density estimate will stay unsmoothed.

In extreme cases it may be that the shape and size of $f(t,\theta)$ and $g(t)$ differ dramatically in areas of high density but the tails of f and g are virtually identical, so a large overall bandwidth would be most suitable. Assume that we initially select h using equation (1), therefore ignoring any prior belief about g, and this chooses a small value. The method of estimating κ given in section 6.4 will result in a small value of p, because $f(t,\theta)$ and $g(t)$ differ so dramatically in D^* that the value of Z^* is still large, despite their similarity in the tails. Thus, when using equation (4) to choose our overall bandwidth, h will remain very small.

Here the introduction of a tail-weighted version of the Z^* function is a feasible solution. We replace Z^* in equation (9) by $Z^*{}_{tail}$, which calculates the difference between expected and observed frequencies only in D^{*C}, where t is such that $h_{t\alpha} > h$ for all t. We take

$$\hat{\kappa} = \left| \frac{n - Z^*{}_{tail}}{Z^*{}_{tail} - 1} \right|,$$

having divided the region D^{*C} into l^* class intervals

$$(\Delta_1, \Delta_2), ..., (\Delta_j, \Delta_{j+1}), ..., (\Delta_{l^*}, \Delta_{l^*+1}).$$

Once again, a suggested guideline for constructing these classes is to create as many as possible, whilst ensuring that $n\hat{r_j}$ is greater than or equal to 5.

Then

$$Z^*{}_{tail} = \frac{1}{l} \sum_{j=1}^{l^*} \frac{(m_j - n\hat{r}_j)^2}{n\hat{r}_j(1 - \hat{r}_j)},$$

with m_j and $\hat{r}_j$ defined as before, as is

$$p = \frac{\hat{\kappa}}{\hat{\kappa} + n}.$$

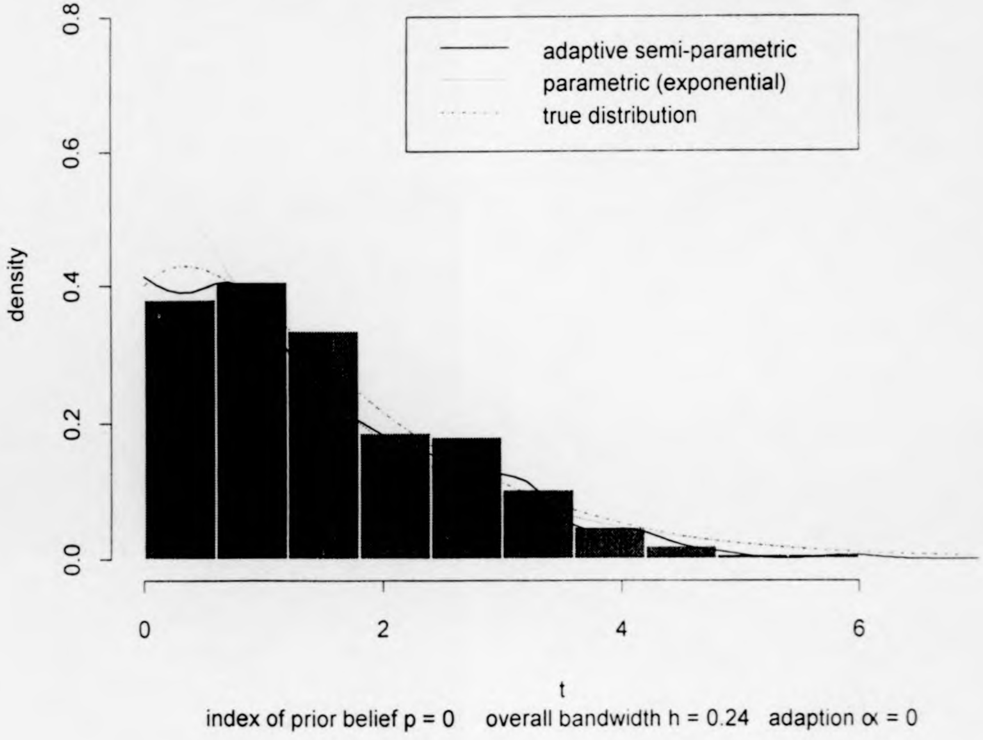
Since we are now only concerned with the difference between $f(t,\theta)$ and $g(t)$ in the tails, when our parametric family is a very good fit in D^{*C} this method will give a larger estimate of κ. This in turn leads to larger values of p, h and α, and the subsequent accuracy of $f(t,\hat{\theta}_{t,\alpha})$ to $g(t)$ in the tails.

6.5.2 An example

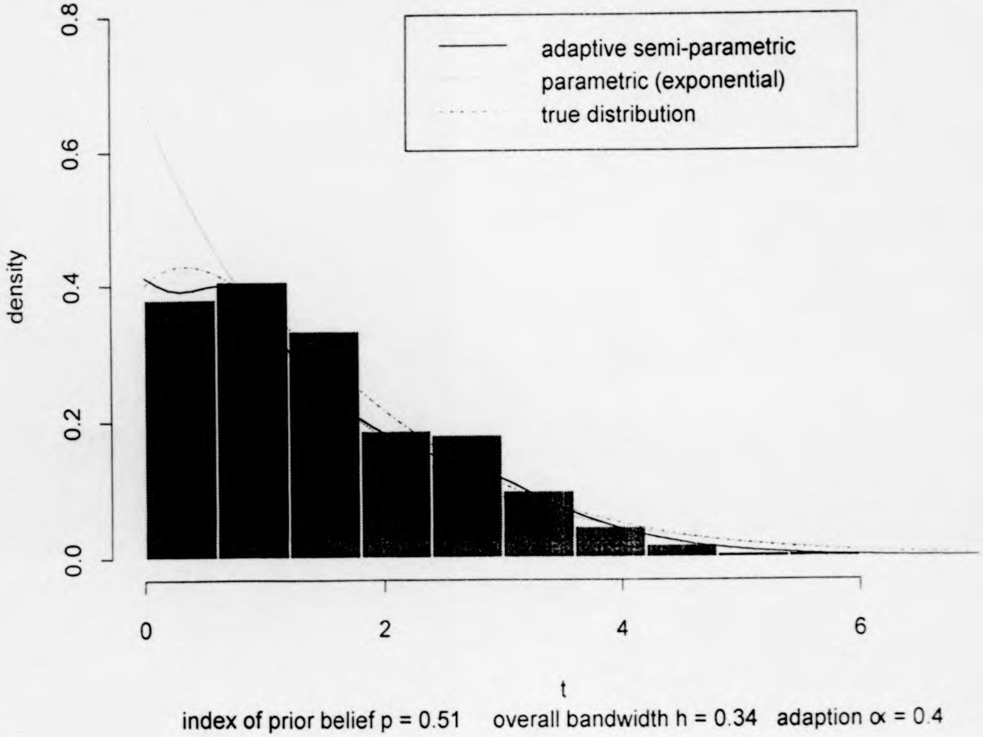
This example considers a simulated data set consisting of a mixture of 100 random points from an exponential(1) distribution and 150 random points from a Gamma[1,2] distribution. A histogram of this data indicates that its density function will be roughly exponential but will flatten off around the mode. It also suggests that the density will be bounded at zero, making a kernel density estimate inherently poor.

Fitting an exponential distribution to the data and using equation (1) to select the overall bandwidth resulted in the very small value of $h = 0.24$ being chosen. This was sufficiently small to give a good, largely non-parametric estimate around the mode, but also left the right tail rather bumpy, as illustrated in plot 6e. Attempting to apply adaption, I found that automatic selection methods (i) and (ii) both chose $\alpha = 0$. Obviously a larger bandwidth would be more suitable, but using the theory of sections 6.2 to 6.4

PLOT 6e: Adaptive semi-parametric estimate without the incorporation of `prior belief'



PLOT 6f: Adaptive semi-parametric estimate with `tail-weighted prior belief'



gave a value of p very close to 0, because of the large difference between the true distribution and our parametric estimate as we approach $t = 0$. The tail-weighted version of Z^* is well suited to this situation; using Z^*_{tail} in the estimation of κ led to a larger value of $p = 0.51$ being chosen, and h was increased by approximately 50 percent when chosen using equation (4). Using method (i) to select α, this larger bandwidth in turn provoked the choice of the small amount of adaption necessary to smooth the left tail of our estimate. The improved adaptive semi-parametric estimate is illustrated in plot 6f. Since in this case we actually know the true distribution and its density function, this is also plotted. Note the small sample size which is largely responsible for the variations of $f(t, \hat{\theta}_{t,\alpha})$ from $g(t)$.

6.5.3 A boundary point-weighted method of estimating κ

Alternatively we may want to increase the value of our automatic choice of h, but feel that the previous suggestions run the risk of moving too far in the opposite direction and producing too large a bandwidth which subsequently provokes the use of too much adaption. This can cause sharp fluctuations in our density estimate near to or at the boundary points, since when α is large we find $f(t, \hat{\theta}_{t,\alpha})$ switching rapidly from approximating a parametric estimate to approximating a kernel estimate.

Another adjustment to the Z^* function estimates κ and chooses p in such a way that the probability of these rapid changes occurring is reduced. We replace equation (12) by a version of Z^* in which extra weighting is given to differences between observed and expected frequencies near the boundary points. The weighting is performed by using an ordinary kernel estimate of the distribution of boundary points. For example, having divided the data into l classes as in the original method, we define

$$Z^*{}_w = \frac{1}{l}\sum_{j=1}^{l}\frac{(n\hat{r}_j - m_j)^2}{n\hat{r}_j(1-\hat{r}_j)}w(j),$$

where t_j is the midpoint of interval (Δ_j, Δ_{j+1}), the set of all q boundary points is $B^* = BP_1, ..., BP_q$, and

$$w(j) = w'(j)\frac{1}{\sum_{j=1}^{l} w'(j)}$$

with

$$w'(j) = \frac{1}{q}\sum_{i=1}^{q}\frac{1}{h^+}K\left(\frac{t_j - BP_i}{h^+}\right). \qquad (14)$$

A Gaussian kernel could be used with bandwidth h^+ chosen as suggested in Silverman (1986), with

$$K(u) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2)$$

and

$$h^{+} = 1.06q^{-\frac{1}{5}}\hat{\sigma_{B*}},$$

where

$$\hat{\sigma_{B*}} = \text{sample standard deviation of the set of boundary points.}$$

When there is only one boundary point, another method of choosing h^{+} will have to be devised. Taking h^{+} equal to one sample standard deviation of the data seems to work well if f is an exponential distribution. As $h^{+} \rightarrow \infty$, then $Z^{*}_{w} \rightarrow Z^{*}$ and choosing a very large value of h^{+} gives results equivalent to those from using the standard method of evaluating $\hat{\kappa}$ and p introduced in section 6.4.

We estimate κ and calculate p with $Z^{*}{}_{w}$ replacing Z^{*} in the modulus of formula (13), such that

$$\hat{\kappa} = \left| \frac{n - Z^{*}_{w}}{Z^{*}_{w} - 1} \right|$$

and

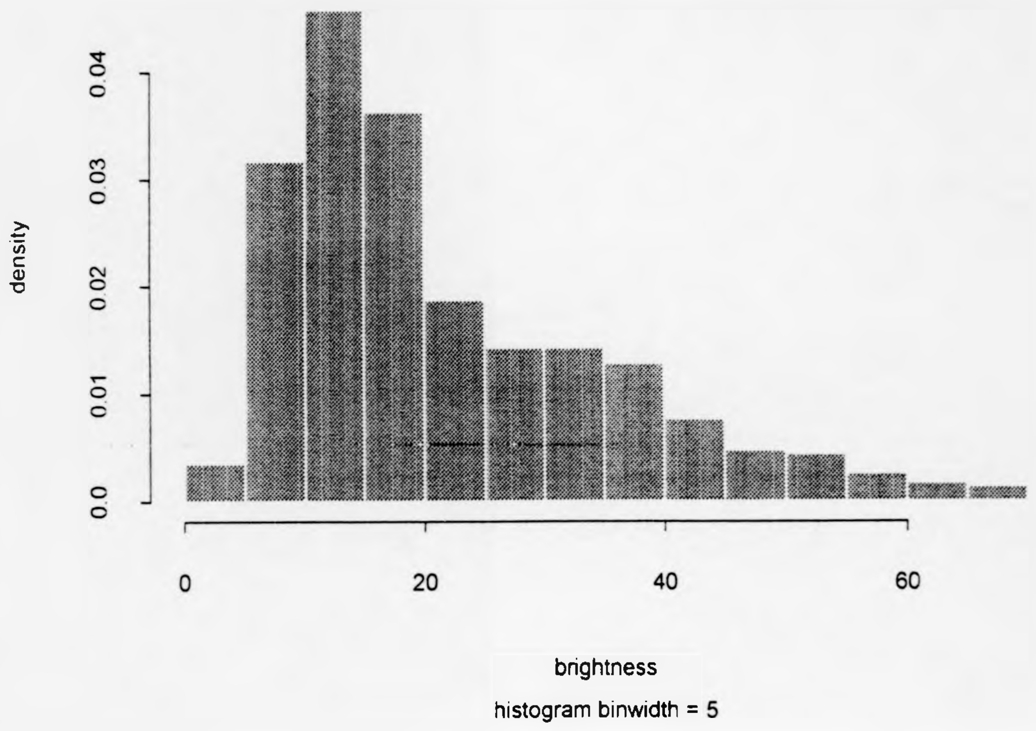
$$p = \frac{\hat{\kappa}}{\hat{\kappa} + n}.$$

This adjustment in the procedure of estimating κ is appropriate when there exists a large local difference between $f(t, \theta)$ and $g(t)$ around the boundary point(s). Since $h_{t\alpha}$ is a continuous function of t, our adaptive semi-parametric estimate $f(t, \bar{\theta}_{t,\alpha})$ will be a smooth curve at the boundary point, but it will possess a very steep gradient here when h and α are large, giving our density estimate in this region an unsatisfactory appearance. The weighted Z_w^* function is likely to detect and prevent problems of this nature.

6.5.4 An example

Example 3 illustrates a situation where this approach for choosing p proved useful. We are attempting to construct a density estimate of brightness reflected from cornfields and detected by satellite. Scott and Factor (1981) apply kernel density estimation to this data set, given in table 3 of their paper. A parametric structure would be preferable, since the non-negativity of the brightness measurements can be accounted for by selecting a parametric family defined only on a positive domain. Copas (1995a) performs semi-parametric density estimation, fitting a Weibull distribution to these data but allowing for local variation.

A histogram of these data, given in plot 6g, confirms a Weibull distribution is indeed a sensible choice, though a purely parametric density estimate underestimates around the mode. Reverting to semi-parametric estimation, PLOT 6g: Normalised histogram of brightness component data indicating the reflectance from cornfields measured by satellite

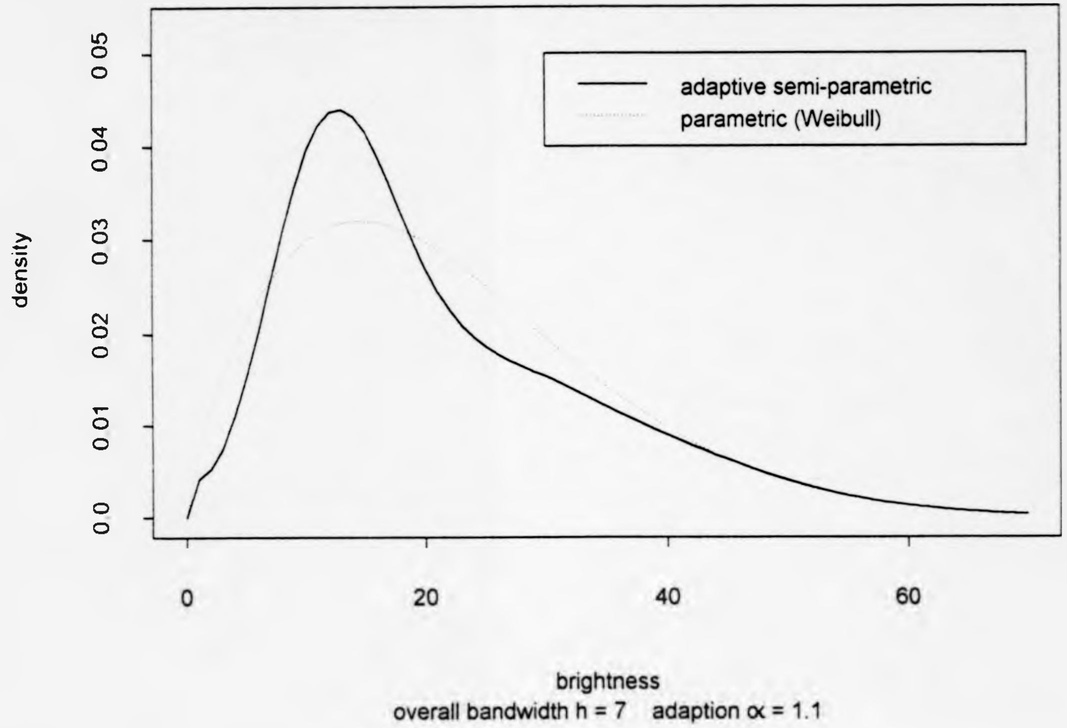


Copas handpicked a bandwidth of $h = 5$; $f(t, \hat{\theta}_{t,\alpha})$ slightly underestimated the true density around the mode, but was a clear improvement over the parametric estimate.

I began by choosing h and α using the automatic methods introduced in chapter 4 and chapter 5, method (iv) respectively. These gave values of $h = 4.1$ and $\alpha = 0$. The resulting density estimate showed improved accuracy around the mode, but now failed to totally smooth out the bumps in the right tail. Handpicking a larger value of h, and again using method (iv) of chapter 5 which chose $\alpha = 1.1$, gave the density estimate shown in plot 6h, in which the increased overall bandwidth h and amount of adaption α has successfully smoothed the right tail. However at the left boundary point, around which $f(t, \hat{\theta})$ and $g(t)$ differ noticeably, the increase in adaption has caused an awkward bump in the density estimate.

A bandwidth larger than 4 appears necessary, but not one so large as to provoke method (iv) into choosing too large a value of α. Instead of resorting to handpicking a smaller value of α than that chosen automatically, we use the boundary point-weighted method to select a more suitable overall bandwidth. We need to select a value of p which will increase h from the value

PLOT 6h: Adaptive semi-parametric density estimate of cornfield reflectance data with no incorporation of prior belief



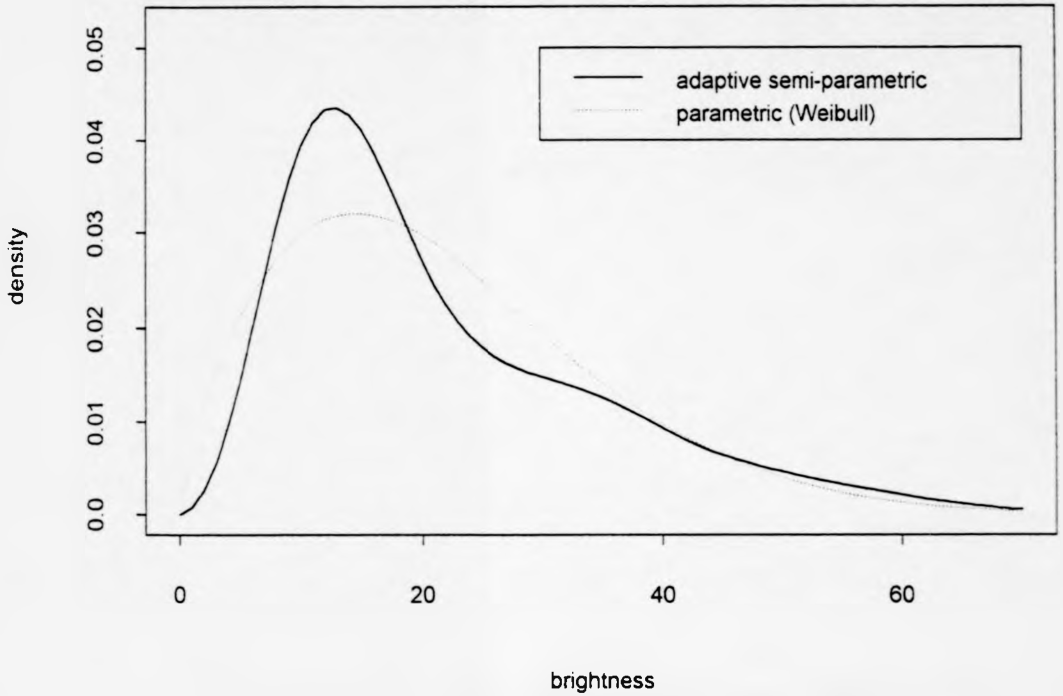
of 4.1 chosen by equation (1), but which takes into account the difference between $f(t,\hat{\theta})$ and $g(t)$ at the boundary points. This is achieved by using Z^*_w as defined above, which leads successively to chosen values of $p = 0.39$, $h = 5.05$ and $\alpha = 0.21$. These produce an excellent density estimate shown in plot 6i. It smooths the right tail satisfactorily, without any problems at the boundary points.

6.6 Estimating κ using a continuous analogy to Z^*

One major problem with the methods outlined in sections 6.4 and 6.5 is that they necessitate the discretisation of the data into l classes. There is no standard procedure for choosing the size of classes, though my suggestion in 6.4 to imitate the standard procedure for the χ^2 statistic worked well in all examples tried, leading to choices of p and h which improved the estimation of the true density. However even within this restriction, slight variations in the choice of $(\Delta_1, \Delta_2), ..., (\Delta_l, \Delta_{l+1})$ caused p to vary by as much as 0.3 in the various examples which I examined.

The following idea sidesteps the discretisation of the data by forming a continuous equivalent to Z^*. Consider that the data have been divided up

PLOT 6i: Adaptive semi-parametric density estimate of cornfield reflectance data with boundary point-weighted prior belief



index of prior belief p = 0.39   overall bandwidth h = 5.05   adaption α = 0.21

into l intervals, of width ϵ, with ϵ being small. If the midpoints of the intervals are defined as $a_1, ..., a_l$, then the vector of probabilities of an observation being in each interval is approximately

$$\epsilon\mathbf{g} = (\epsilon g_1, ..., \epsilon g_l)^T,$$

where $g_i = g(a_i)$, the true density at the ith midpoint. This approximation improves as $l \longrightarrow \infty$, causing $\epsilon \longrightarrow 0$. We suppose that vector $\epsilon\mathbf{g}$ is distributed Dirichlet given prior distribution $\epsilon\mathbf{f}$, where $f_i = f(a_i, \theta)$ is the density function of our chosen parametric family at the ith midpoint, such that

$$\epsilon\mathbf{f} = (\epsilon f_1, ..., \epsilon f_l)^T$$

and

$$\epsilon\mathbf{g}|\epsilon\mathbf{f} \sim \text{Dirichlet}(\kappa, \epsilon\mathbf{f}).$$

The number of observations in each interval is given by vector $\mathbf{m}$, where

$$\mathbf{m}|\epsilon\mathbf{g} \sim \text{Multinomial}(n, \epsilon\mathbf{g})$$

As in the discrete formulation of section 6.4, we use conditional expectations to calculate

$$E\left(m_i\right) = E\left(E(m_i|g_i)\right) = E\left(n\epsilon g_i\right) = n\epsilon f_i, \tag{15}$$

and following the same procedure as in equations (9) to (11), we get

$$Var\,(m_i) = Var\,(E(m_i|g_i)) + E\,(V(m_i|g_i))$$

$$= n\epsilon f_i(1-\epsilon f_i)\left(1+\frac{n-1}{\kappa+1}\right). \tag{16}$$

The covariances can also be evaluated, using

$$Cov\,(m_i,m_j) = E\,(Cov(m_i,m_j|g_i,g_j)) + Cov\,(E(m_i|g_i),E(m_j|g_j))$$

$$= nE(\epsilon g_i\epsilon g_j) + n^2Cov(\epsilon g_i,\epsilon g_j)$$

$$= -n\epsilon^2 f_if_j + (n^2-n)Cov(\epsilon g_i,\epsilon g_j)$$

$$= -n\epsilon^2 f_if_j\left(1+\frac{n-1}{\kappa+1}\right). \tag{17}$$

We now smooth the m_i's, defining

$$M_i = \sum_{j=1}^{l} m_jw_{ij},$$

a weighted sum of the number of observations in the l intervals, with w_{ij} chosen as

$$w_{ij} = \frac{1}{nh^\circ}\phi\left(\frac{a_i-a_j}{h^\circ}\right).$$

Calculating the expectations and variances of M_i using the expectation, variance and covariance calculations of equations (15) to (17), we find that

$$E(M_i) = \sum_{j=1}^{l} n\epsilon f_jw_{ij} \tag{18}$$

and

$$\begin{aligned} Var(M_i) &= \sum_{j=1}^{l} Var(m_j w_{ij}) + \sum_{j,k:j\neq k}^{l} Cov(m_j w_{ij}, m_k w_{ik}) \\ &= \left(1+\frac{n-1}{\kappa+1}\right)\left(\sum_{j=1}^{l} w_{ij}^2 n\epsilon f_i(1-\epsilon f_i) - \sum_{j,k:j\neq k}^{l} w_{ij}w_{ik}n\epsilon^2 f_j f_k\right) \\ &= n\left(1+\frac{n-1}{\kappa+1}\right)\left(\sum_{j=1}^{l} w_{ij}^2\epsilon f_i - \left(\sum_{j=1}^{l} w_{ij}\epsilon f_j, \theta)\right)^2\right). \end{aligned} \tag{19}$$

If we now take the limit as $\epsilon \to 0$, implying that $l \to \infty$, then

$$M_i = \frac{1}{nh^\circ}\sum_{j=1}^{l}\phi\left(\frac{a_i - a_j}{h^\circ}\right) m_j \to \hat{g}^\circ(a_i),$$

the ordinary kernel estimate of true density $g(t)$ at target point $t = a_i$, using a Gaussian kernel with bandwidth h°.

Applying this limiting procedure to our expectations and variances we find that

$$E(M_i) = \sum_{j=1}^{l}\frac{1}{h^\circ}\phi\left(\frac{a_i - a_j}{h^\circ}\right)\epsilon f_j \to \int_u \frac{1}{h^\circ}\phi\left(\frac{a_i - u}{h^\circ}\right) f(u,\theta)du = E_f\left(\hat{g}^\circ(a_i)\right), \tag{20}$$

and similarly the variance converges such that

$$Var(M_i) = \frac{1}{n}\left(1+\frac{n-1}{\kappa+1}\right)$$

$$\left(\frac{1}{h^{\circ 2}}\sum_{j=1}^{l}\phi^2\left(\frac{a_i - a_j}{h^\circ}\right)\epsilon f_j - \frac{1}{h^{\circ 2}}\left(\sum_{j=1}^{l}\phi\left(\frac{a_i - a_j}{h^\circ}\right)\epsilon f_j\right)^2\right)$$

$$\rightarrow \frac{1}{n}\left(1+\frac{n-1}{\kappa+1}\right)$$

$$\left(\frac{1}{h^\circ}\int_u \frac{1}{h^\circ}\phi^2\left(\frac{a_i-u}{h^\circ}\right) f(u,\theta)du - \left(\int_u \frac{1}{h^\circ}\phi\left(\frac{a_i-u}{h^\circ}\right) f(u,\theta)du\right)^2\right). \quad (21)$$

Under the limit $\epsilon \rightarrow 0$, the variance of M_i can also be written as

$$Var(M_i) = E\left(\hat{g}^\circ(a_i) - E_f(\hat{g}^\circ(a_i))\right)^2,$$

so rearranging as before and writing in terms of the target points t, we have

$$\left(1+\frac{n-1}{\kappa+1}\right) =$$

$$\left(\frac{E_f\left(\hat{g}^\circ(t) - E_f(\hat{g}^\circ(t))\right)^2}{\frac{1}{n}\left(\frac{1}{h^\circ}\int_u \frac{1}{h^\circ}\phi^2\left(\frac{u-t}{h^\circ}\right) f(u,\theta)du - \left(\int_u \frac{1}{h^\circ}\phi\left(\frac{u-t}{h^\circ}\right) f(u,\theta)du\right)^2\right)}\right).$$

We now estimate the right-hand side of this equation by

$$Z_c^* =$$

$$\int_t \left(\frac{\left(\hat{g}^\circ(t) - \int_u \frac{1}{h^\circ}\phi\left(\frac{u-t}{h^\circ}\right) f(u,\hat{\theta})du\right)^2}{\frac{1}{n}\left(\frac{1}{h^\circ}\int_u \frac{1}{h^\circ}\phi^2\left(\frac{u-t}{h^\circ}\right) f(u,\hat{\theta})du - \left(\int_u \frac{1}{h^\circ}\phi\left(\frac{u-t}{h^\circ}\right) f(u,\hat{\theta})du\right)^2\right)}\right) f(t,\hat{\theta})dt. \quad (22)$$

Bandwidth h° is chosen using the simple plug-in formula suggested in Silverman (1986). The integral over t is evaluated using numerical integration over a fine grid of t's. Parameter κ is now estimated by

$$\hat{\kappa} = \left|\frac{n - Z_c^*}{Z_c^* - 1}\right|, \quad (23)$$

and p is again taken as

$$p = \frac{\hat{\kappa}}{\hat{\kappa} + n}.$$

The statistic Z_c^* is a continuous analogy to the Z^* statistic defined in equation (12).

When calculating $\hat{\kappa}$ and thus p from equations (19) to (22), I rapidly performed the necessary numerical integration by Simpson's Rule using a computer.

I have only been able to apply this method when the integrals with respect to u in equation (22) have been analytically calculable in terms of t. This is the case when we take parametric family f to be Normal or exponential. If it is possible to use this method, it has much to recommend it, since it removes the need to choose class intervals. Different choices of class intervals will give different results in the discrete method outlined in section 6.4. This makes it hard to compare results from the continuous and discrete methods. In many cases the value of p extracted by the continuous method was close to that chosen by the corresponding discrete method with one particular division of the data, but differed noticeably from the value found via another division.

It is also possible to build weighting into the continuous method of estimating κ, as was done for the discrete method. A continuous analogy exists for the discrete tail-weighted method of subsection 6.5.1. Rather than integrating with respect to t over the domain of f, we could integrate only over the interval $t \in D^{*C}$, where D^* is as defined in Chapter 3. An analogy to the boundary point-weighted version of subsection 6.5.3 could be constructed by inserting an appropriate weight function into equation (22), such as a kernel density estimate of the boundary points themselves.

6.7 Comment

It is worth reiterating that the methods outlined above are more in the nature of sketched suggestions than the final word on choosing our overall bandwidth h for use in the adaptive semi-parametric density estimation procedure.

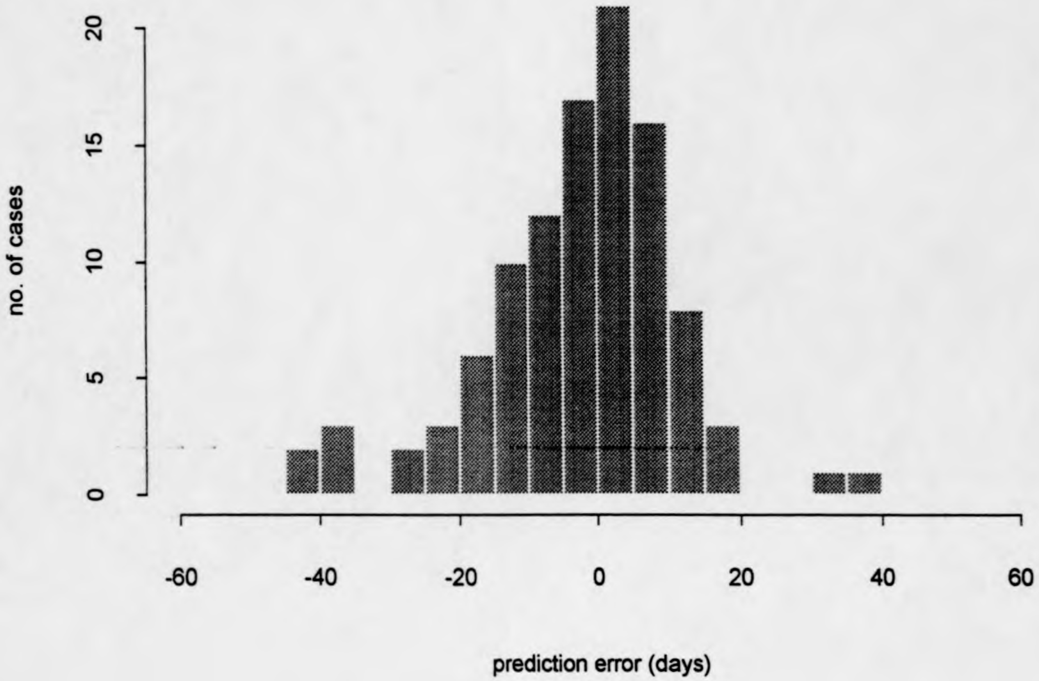
7 Real data examples and Conclusions

7.1 Birth prediction error data

This data set was taken from a British doctor's *Medical Diary and Visiting List* dating back to the year 1916. The "Obstetric Engagements" section contained records of expected and actual birth dates for 127 women who had passed through pregnancy under the doctor's care. I have converted this data into a record of **birth date prediction error**, defined as **the difference in days between the date when a birth was expected and when it actually took place**. Premature births are recorded as negative values. The 22 patients for whom at least one of the dates was not recorded were ignored, leaving a sample size of 105.

The medical diary was discovered by Dr Stephen Senn in the mid-seventies and an initial analysis of various data sets extracted from it are given in Senn (1979). More recently he has concentrated in particular on this obstetric data, with Senn (1995),"A General Practitioner's Obstetric Diary", currently submitted for publication. This paper suggests that estimating the density of prediction error would be an interesting avenue to explore, a histogram of the prediction error data, shown in plot 7a, illustrating why. A Normal fit

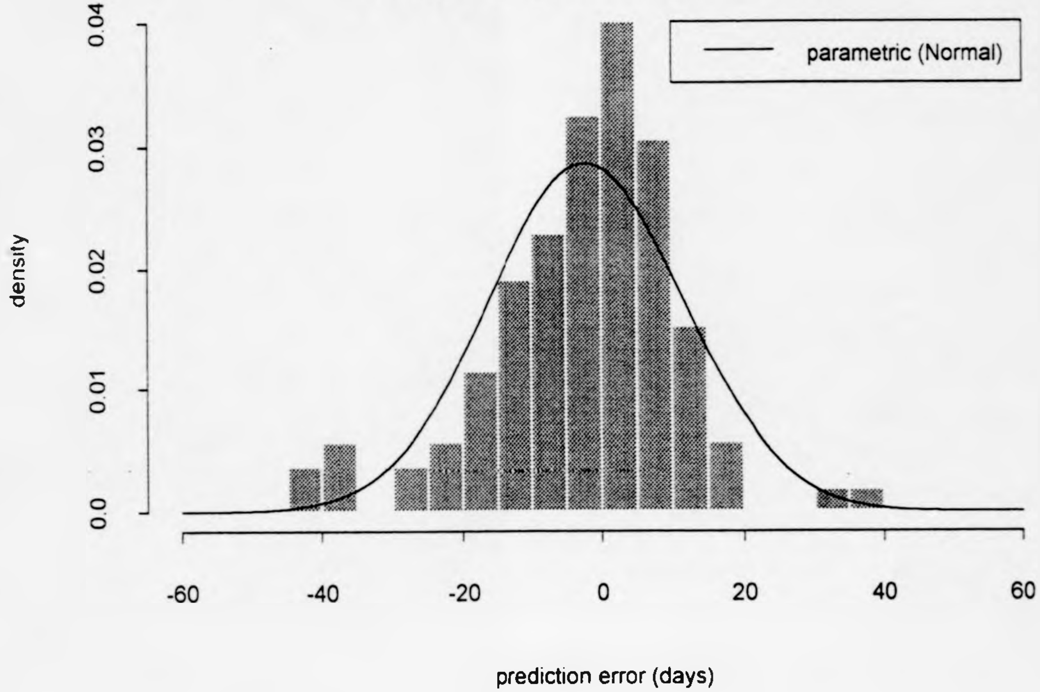
PLOT 7a: Histogram of Birth Date Prediction Error



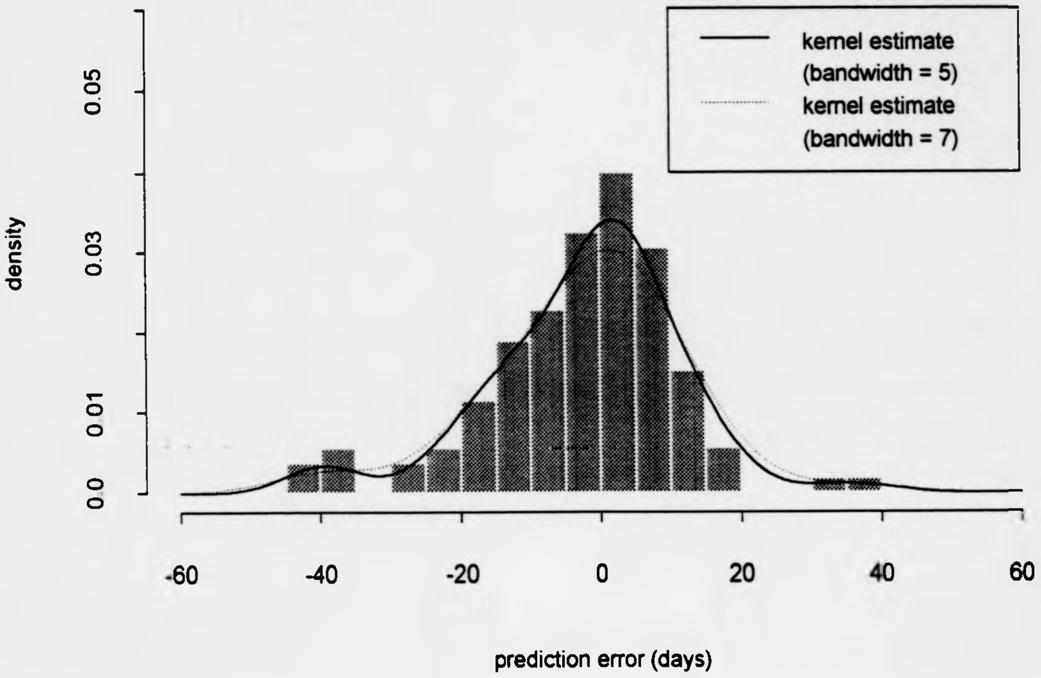
to the data seems possible in the tails which appear to be of similar shape. However this is ineffective in capturing the behaviour around the mode, where the histogram suggests that the true density takes slightly larger values than those of the parametric estimate shown in plot 7b. The data exhibits a small degree of skewness which our parametric estimate also fails to pick out. Meanwhile, non-parametric kernel estimation struggles because of the small sample size, with the large bandwidth necessary to smooth the tails leading to oversmoothing at the mode. Plot 7c illustrates this dilemma, with two kernel estimates shown. The estimate which used the smaller bandwidth of 5, chosen by the plug-in method from Silverman (1986), has a noticeable bump in the left tail, while the use of a larger handpicked bandwidth of 7 causes some underestimation of the density in high density regions. The smoothness and accuracy of our estimate in the tails are especially important with particular dataset; prediction interests are likely to concern the probabilities of births being overdue or premature by a certain number of days, and calculating these probabilities requires accurate estimation in tail areas.

Our adaptive semi-parametric method would appear to be ideal for use in this situation, enabling us to leave the tails well smoothed whilst picking out local deviations from the parametric Normal fit in areas of high density.

PLOT 7b: Normalised histogram of prediction error and parametric density estimate



PLOT 7c: Normalised histogram of prediction error data and non-parametric ordinary kernel density estimates

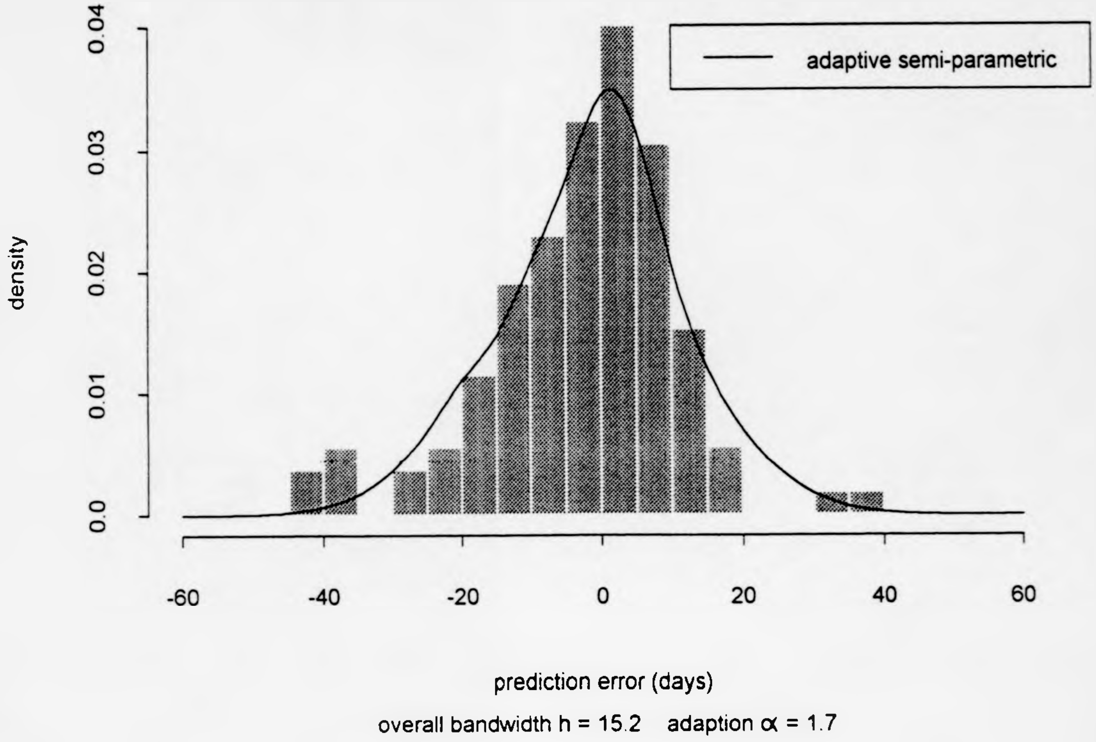


We used the automatic method of chapter 4, equation (16) to select our bandwidth, along with method (iv) of chapter 5 for automatically selecting α. These gave values of $h = 15.2$ and $\alpha = 1.7$ respectively, resulting in the adaptive semi-parametric density estimate pictured in plot 7d. While there may still be some slight underestimation of the density at the mode, it is a clear improvement on both the parametric and kernel density estimates, picking out the skewness and the high modal density. Though the data set is relatively small, the automatic selection methods for h and α appear to have performed well again. Slightly different values of these parameters produced estimates which were no better than that depicted in plot 7d.

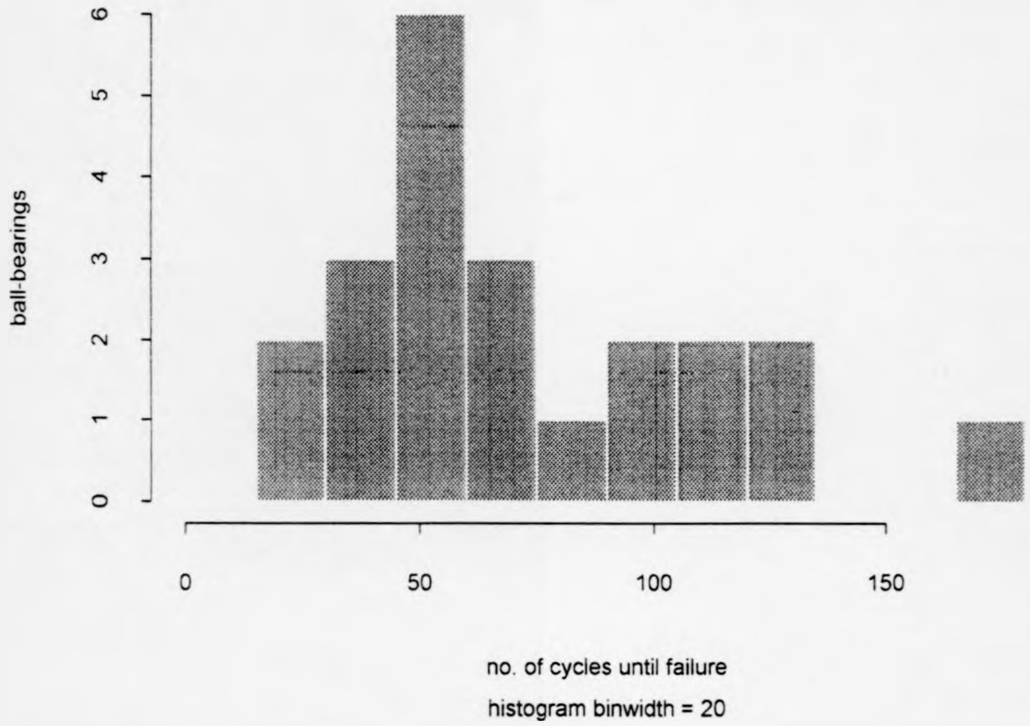
7.2 A small data set example

This next data set can be found in "Small Data Sets", Hand et al (1995). It consists of 'cycles until failure' of deep-groove ball-bearings, and was originally published by Lieblein and Zelen (1956), who argued that these data could be modelled as a Weibull distribution. Estimation of failure time would obviously be a major part of any analysis of these data. A histogram of the data shown in plot 7e confirms the plausibility of a Weibull fit, though the

PLOT 7d: Normalised histogram of prediction error with adaptive semi-parametric density estimate



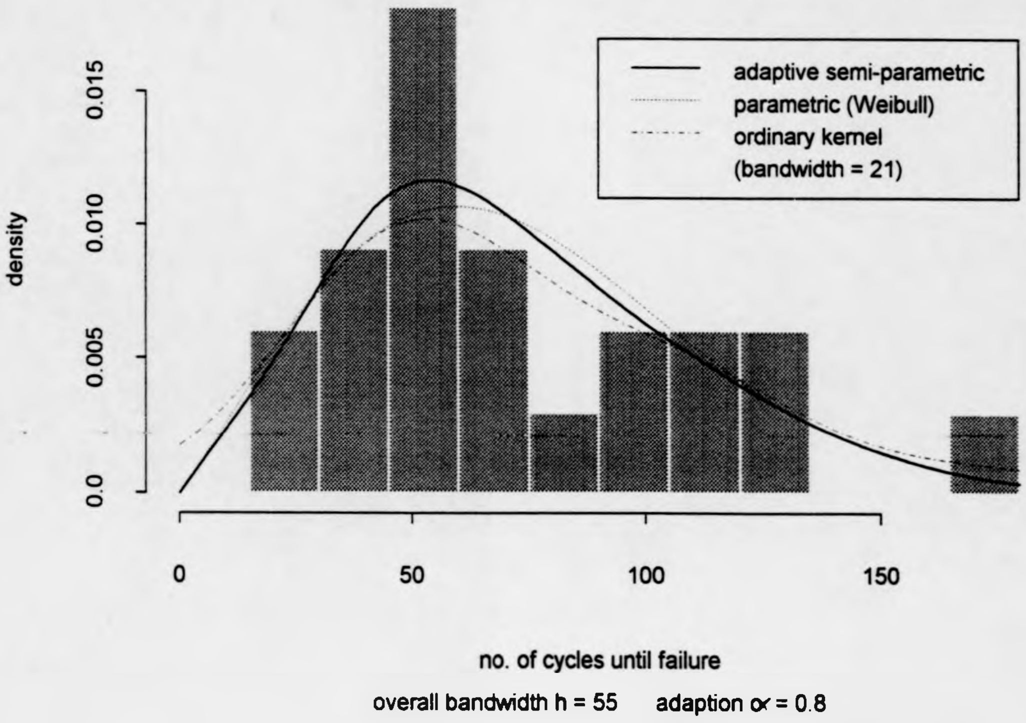
PLOT 7e: Histogram of ball-bearing fatigue data



density around the mode appears to be slightly underestimated. A non-parametric kernel estimate, with bandwidth chosen by the plug-in method of Silverman (1986), struggled to deal with the bounded left tail of the density function and gave a positive density estimate 'behind' zero. An adaptive semi-parametric density estimate retains the bound at zero due to its parametric structure, and should cope better around the mode as it will be locally influenced by the data.

This is a **very** small data set, containing only 22 observations, and the use of semi-parametric methods is perhaps not ideal. Comparisons with other methods are foolhardy though we can show that $f(t, theta_{t,\alpha})$ is at least a reasonable density estimate. However, it is a significant test of our automatic methods for selecting the parameters h and α used in the construction of our adaptive semi-parametric estimate. These selection procedures were based on small h and large n approximations, but they again perform very satisfactorily, with the adaptive semi-parametric estimate pictured in plot 7f giving a slightly higher density estimate at the mode. We would not want our estimate to follow the shape of the normalised histogram too tightly since it is such a small data set, with plenty of scope for sampling error. Our selections of h and α, again made using equation (16) of chapter 4 and method (iv)

PLOT 7f: Normalised histogram of fatigue data with parametric, ordinary kernel and adaptive semi-parametric density estimates



from chapter 5 respectively, appear to have allowed exactly the right amount of local influence given the small sample size.

7.3 Problems and avenues for further research

Local likelihood related semi-parametric density estimation is a very new field of study, and as such there exist many opportunities for further research based upon the ideas of this thesis and the other literature touched upon in chapter 1. Throughout the writing of this thesis several problems and possibilities have emerged, for which there has not been sufficient time to solve or explore further.

7.3.1 Introducing adaption into other semi-parametric methods

In chapter 1, two alternative semi-parametric methods based upon maximising different local likelihood functions were outlined, one of which (that of Hjort and Jones) has already been published. It too provides a continuum between parametric and non-parametric estimation controlled by a smoothing parameter. There is no reason why this smoothing value should not vary with the target point at which we are estimating, as $h_{t,\alpha}$ does in the adaptive semi-parametric method considered in this thesis.. This could be achieved simply by replacing the weight function in equation (7) of section 1.3 by a local kernel function. Locally varying bandwidth h_t could again be taken as $h_{t,\alpha}$ given in equation (4) of chapter 3. The different structure of Hjort and Jones's local likelihood function would require new approximations and different methods of selecting h and α. Similarly, the practice of varying the bandwidth locally could be applied in the local likelihood function developed by Eguchi, though the poor performance of his technique as the bandwidth decreases makes this a less attractive proposition.

7.3.2 The formulation of $h_{t,\alpha}$ in the local kernel function

The local kernel function is used in the weighting of our local likelihood function, but can be used as a density estimation method in its own right. In the past, one drawback has been the lack of a straightforward procedure of selecting the local bandwidth h_t. Previously the accepted method has been to choose h_t for each target point to minimise the small h approximation of the MSE at t. This process involves estimating the second derivative of the true distribution, and breaks down when this estimate is equal to 0. My construction of $h_t = h_{t,\alpha}$ avoids these problems, and the two parameter set up allows for greater flexibility in how and from where our bandwidth varies.

7.3.3 Varying α with respect to location

A further extension to the ideas of chapter 3 would be to vary α with respect to location. One advantage of this would be to enable α to decrease as we approach any boundary points. This would be useful in cases where the parametric and non-parametric density estimates differ substantially at one or more boundary points, but where we would still like to select a large overall bandwidth and have large α values in other areas. It would be an alternative to using the ideas of subsection 6.5.3, which reduce h and thus our automatic selection of α, so avoiding a sudden jump in our adaptive semi-parametric estimate from approximating $\hat{g}_{L,\alpha}(t)$ to approximating $f(t,\theta)$. A gradual reduction in α as we approach the boundary points, thus making the convergence of $h_{t,\alpha}$ to h from either side of the boundary point less rapid, would ensure smoother estimates in this region. How exactly to formulate a location-variable α value is an open question.

One possibility is to use a similar structure to that of our local bandwidth $h_{t,\alpha}$. We could locally vary the amount of adaption from a baseline or overall adaption value α, such that at target point t, the amount of adaption applied is defined as

$$\alpha_t = a(t)\alpha.$$

Function $a(t)$ controls how much and in which direction our local adaption α_t differs from α.

Bearing in mind the motivation described above, we could for example reduce the amount of adaption used as we approach the boundary points, by choosing

$$a(t) = \left(1 - \frac{f(t,\hat{\theta})}{\lambda}\right)^2 = (1 - \Upsilon(t))^2.$$

Then our local bandwidth at t would become

$$h_{t,\alpha_t} = h\Upsilon(t)^{-\alpha_t} = h\Upsilon(t)^{-\alpha(1-\Upsilon(t))^2}.$$

As $t \rightarrow t^*$, where t^* is a boundary point, then

$$a(t) \rightarrow 0 \Rightarrow \alpha_t \rightarrow 0 \Rightarrow h_{t,\alpha_t} \rightarrow h.$$

The behaviour at the boundary points will not be affected; $\Upsilon(t) = 1$ at $t = t^*$, so adaption had no effect there anyway. If the domain of the density function defining f is unbounded, $\alpha_t \rightarrow \alpha$ as $|t| \rightarrow \infty$. When choosing α, we should bear this in mind, as well as the fact that α_t will only exceed α when $f(t,\hat{\theta})$ gets very large.

7.3.4 A plug-in selection method for h and α

Chapters 4 and 5 gave several ideas on how to select h and α. However all involved the use of computer-based minimisation procedures and in some cases numerical integration. While they were very quick and easy to compute, the option of a neater plug-in formulae is desirable. Something similar to that given in Silverman for choosing the bandwidth in ordinary kernel estimation would be ideal, if just to give a rough guide as to what values of h and α we should consider implementing, without the use of a computer.

7.3.5 Dealing with awkwardly shaped distributions

Finally a continual problem in evolving methods for parameter selection has been dealing with cases where the data is from a distribution which is bounded, contains discontinuities and differs drastically in shape to the Normal distribution. Such cases cause problems due to the need for preliminary estimation of $g(t)$ and its first two derivatives, for which we use kernel estimation. Our choice of a Gaussian kernel results in poor preliminary estimates of extremely non-Normal density functions; bounded and discontinuous functions will always pose difficulties for the kernel method.

An example of a distribution which exhibits all of these problematic characteristics is the exponential. Apart from producing density estimates which integrate to values noticeably less than one, the adaptive semi-parametric method is ideal for dealing with exponentially distributed data since it is able, through the incorporation and imposition of a parametric family, to retain shape and boundaries, unlike non-parametric methods such as ordinary kernel estimation. See, for example, the frequently used example of the deer line transect data, or the remand data example from Copas (1995a). However the methods suggested for selecting h and α require preliminary rough estimates of the true density and its derivatives, which on occasions have been so poor as to cause these methods to break down. New techniques to deal with such cases would be helpful.

REFERENCES

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

Abramson,I. (1982) Arbitrariness of the pilot estimator in adaptive kernel methods. *Journal of Multivariate Analysis* **12**, 302–307.

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

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

Buckland,S.T. (1992) Fitting density functions with polynomials. *Applied Statistics* **41**, 63–76.

Burnham,K.P., Anderson,D.R. and Laake,J.L. (1980) Estimation of density from line transect sampling of biological populations. Supplement to *The Journal of Wildlife Management* **44**, No. 72.

Clarke,B.S. and Barron,A.R. (1990) Information-theoretic asymptotics of Bayes methods. *IEEE Trans.Inform* **36**, No. 3, 453–471.

Cleveland,W.S. (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829–836.

Copas,J.B. (1995a) Local likelihood based on kernel censoring. *Journal of the Royal Statistical Society Series B* **57**, 221–235.

Copas,J.B. (1995b) Semi-parametric density estimation by likelihood. Research Report 262, Statistics Department, University of Warwick (submitted to *Journal of the Royal Statistical Society Series B*).

Copas,J.B. and Stride,C.B. (1995) Fitting a Normal distribution with local influence. Research Report 269, Statistics Department, University of Warwick (submitted to *Journal of the Royal Statistical Society Series B*).

Cox, D.R. (1972) Regression models and life tables. *Journal of the Royal Statistical Society Series B* **34**, 187–202.

Duin,R.P.W. (1976) On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Trans.Comput* **C-25**, 1175–1179.

Fan,J. (1992) Design-adaptive nonparametric regression. *Journal of the American Statistical Association* **87**, 998–1004.

Fan,J. (1993) Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics* **21**, 196–216.

Fan,J. and Gibjels,I. (1992) Variable bandwidth and local linear regression smoothers. *Annals of Statistics* **20**, 2008–2036.

Fan,J. and Marron,J.S. (1992) Best possible constant for bandwidth selection. *Annals of Statistics* **20**, 2057–2070.

Hall,P. (1987) On Kullback-Leibler loss and density estimation. *Annals of Statistics* **15**, 1491–1519.

Hall,P. (1990) On the bias of variable bandwidth curve estimators. *Biometrika* **77**, 529–535.

Hall,P. (1992) On global properties of variable bandwidth density estimators. *Annals of Statistics* **20**, 762–78.

Hall,P., Sheather,S.J., Jones,M.C. and Marron J.S. (1991) On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* **78**, 263–269.

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

Hand,D.J., Daly,F., Lunn,A.D., McConway,K.J. and Ostrowski,E. (1995) *A Handbook of Small Data Sets.* Chapman and Hall, London.

Hastie,T. and Tibshirani,R. (1990) *Generalised Additive Models.* Chapman and Hall, London.

Hjort,N.L. and Jones,M.C. (1994) Locally parametric nonparametric density estimation. Research Report 3, Institute of Mathematics, University of Oslo.

Jones,M.C. (1990) Variable kernel density estimates and variable kernel density estimates. *Australian Journal of Statistics* **30**, 361–371.

Lieblein,J. and Zelen,M. (1956) Statistical investigation of the fatigue-life of deep groove ball-bearings. *Journal of Research, National Bureau of Standards* **57**, 273–316.

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

Olkin,I. and Speigelman,C.H. (1987) A semiparametric approach to density estimation. *Journal of the American Statistical Association* **82**, 858–865.

Rudemo,M. (1982) Empirical choice of histograms and kernel density estimators. *Scandanavian Journal of Statistics* **9** 65–78.

Schucany,W.R. (1989) Locally optimal window widths for kernel density estimation with large samples. *Statistics and Probability Letters* **7**, 401–405.

Scott, D.W. and Factor, L.E. (1981) Monte-Carlo study of three data-based nonparametric probability density estimators. *Journal of the American Statistical Association* **76**, 9–15.

Senn,S.J. (1979) A sixty year old medical record. *Medical Record and Health Care Information Journal* **20**, 528-531.

Senn,S.J. (1995) A general practitioners obstetric diary. Manuscript (submitted to *Data and Statistics*).

Sheather,S.J. (1986) An improved data-based algorithm for choosing the window width when estimating the density at a point. *Computational Statistics and Data Analysis* **4**, 61–65.

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

Tibshirani,R. and Hastie,T. (1987) Local likelihood estimation. *Journal of the American Statistical Association* **82**, 559–567.

This thesis was produced according to the guidelines laid down by The University of Warwick Graduate School. CBS 15/9/95