ISSN 1440-771X
ISBN 0 7326 1053 2

MONASH UNIVERSITY

Bandwidth Selection for Kernel Conditional Density Estimation

David M. Bashtannyk and Rob J. Hyndman

Working Paper 16/98
October 1998

DEPARTMENT OF ECONOMETRICS
AND BUSINESS STATISTICS

Bandwidth selection for kernel conditional density estimation

David M Bashtannyk and Rob J Hyndman[1]

14 October 1998

Abstract: We consider bandwidth selection for the kernel estimator of conditional density with one explanatory variable. Several bandwidth selection methods are derived ranging from fast rules-of-thumb which assume the underlying densities are known to relatively slow procedures which use the bootstrap. The methods are compared and a practical bandwidth selection strategy which combines the methods is proposed. The methods are compared using two simulation studies and a real data set.

Keywords: density estimation; kernel smoothing; conditioning; bandwidth selection.

1 Introduction

Kernel conditional density estimation was first considered by Rosenblatt (1969) who studied the problem of estimating the density of Y conditional on $X = x$ where X is univariate and random. If $g(x, y)$ denotes the joint density of (X, Y) and $h(x)$ denotes the marginal density of X, then the conditional density of $Y\,|(X = x)$ is given by $f(y \mid x) = g(x, y)/h(x)$.

Rosenblatt proposed the following kernel estimator of f:

$$\hat{f}(y \mid x) = \frac{\frac{1}{nab}\sum_{j=1}^{n} K\left(\frac{\|x-X_j\|_x}{a}\right) K\left(\frac{\|y-Y_j\|_y}{b}\right)}{\frac{1}{na}\sum_{j=1}^{n} K\left(\frac{\|x-X_j\|_x}{a}\right)} \tag{1.1}$$

where $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ is a sample of independent observations from the distribution of (X, Y) and $\|\cdot\|_x$ and $\|\cdot\|_y$ are distance metrics on the spaces of X and Y respectively.

The kernel function, $K(u)$, is assumed to be a real, integrable, non-negative, even function on $\Re$ concentrated at the origin such that

$$\int_{\mathbf{R}} K(u)du = 1, \quad \int_{\mathbf{R}} uK(u)du = 0 \quad \text{and} \quad \sigma_K^2 = \int_{\mathbf{R}} u^2 K(u)du < \infty. \tag{1.2}$$

Popular choices for $K(u)$ are defined in terms of univariate and unimodal probability density functions.

[1]Department of Econometrics and Business Statistics, Monash University, Clayton 3168, Australia. Correspondence concerning this article should be directed to Rob Hyndman (Email: Rob.Hyndman@monash.edu.au)

The problem of conditional density estimation appears to have lain free of scrutiny until it was revisited recently and some improved estimators were proposed.

Hyndman, Bashtannyk & Grunwald (1996) give the bias, variance, MSE and convergence properties of $\hat{f}(y \mid x)$ and proposed a modified kernel estimator with smaller MSE than the standard estimator in some commonly occurring situations. Fan, Yao & Tong (1996) proposed an alternative conditional density estimator by generalizing Rosenblatt's estimator using local polynomial techniques. Stone (1994) followed a different path and considered using tensor products of polynomial splines to obtain conditional log density estimates.

In this paper we consider the problem of bandwidth selection for Rosenblatt's original estimator. We also comment on how to extend the ideas presented here to the improved estimators introduced later.

We shall rewrite (1.1) as

$$\hat{f}(y \mid x) = \frac{1}{b} \sum_{j=1}^{n} w_j(x) K\left(\frac{\|y - Y_j\|_y}{b}\right) \tag{1.3}$$

$$\text{where} \qquad w_j(x) = \frac{K\left(\frac{\|x - X_j\|}{a}\right)}{\sum_{i=1}^{n} K\left(\frac{\|x - X_i\|}{a}\right)}$$

The parameters a and b control the degree of smoothing applied to the density estimate; a controls the smoothness between conditional densities in the x direction and b controls the smoothness of each conditional density in the y direction. The selection of a and b has a critical role in determining the performance of the kernel conditional density estimate.

Figure 1 shows graphically how the kernel conditional density estimate is constructed. For simplicity we have used 20 observations, although a much higher number of observations is required for meaningful conditional density estimation.

Figure 1(a) shows kernel functions with bandwidth b, centered at the observations. The conditioning $X = x$ is carried out by another kernel function in the X-space. This second kernel function has bandwidth a and is centered at the conditioning value $x = x_0$. (This kernel is normalized so that the total weights sum to one.) The kernel function chosen for this illustration has bounded support and observations outside the window width a carry zero weight. The shaded region shows those observations which have non-zero weight.

In Figure 1(b) the conditional density estimate at $X = x_0$ is shown. This was obtained by summing the n kernel functions in Y-space, weighted by the kernel function in X-space.

Our approach in bandwidth selection will be to minimize the weighted integrated mean square error function (IMSE), defined as

$$\text{IMSE}(a, b; \hat{f}, f) = \iint \mathsf{E}\left\{\hat{f}(y \mid x) - f(y \mid x)\right\}^2 h(x)\, dx\, dy. \tag{1.4}$$

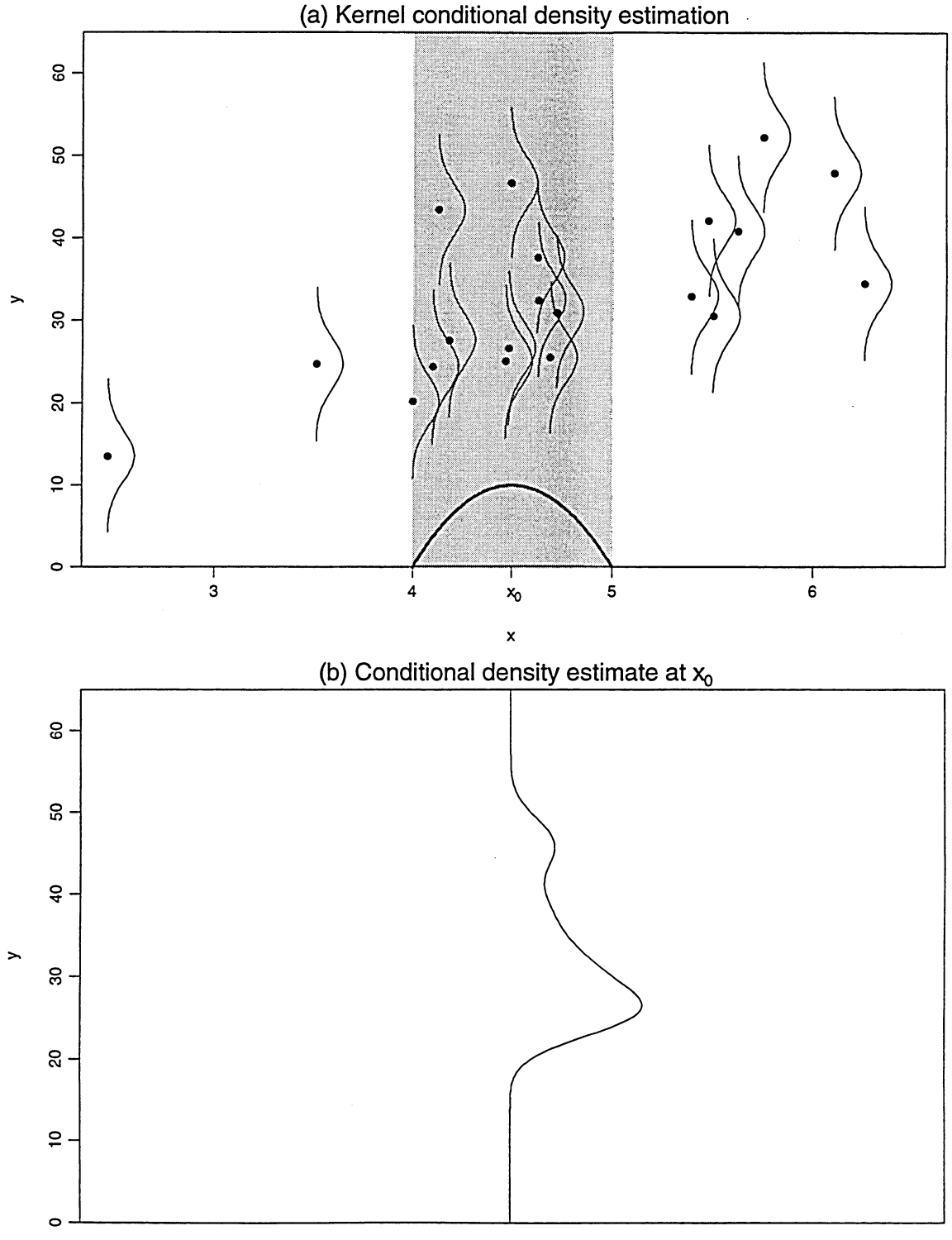

Figure 1: *Construction of the kernel conditional density estimate* $f(y \mid x_0)$ *at the conditioning value* $X = x_0$. *The shaded region shows the observations which receive non-zero weight. The weight function is shown as the heavy line in the top plot.*

Weighting the IMSE by the marginal density $h(x)$ places more emphasis on the regions that have more data and it also eases the computational difficulty.

We also define the integrated square error function (ISE) as

$$\text{ISE}(a,b;\hat{f},f) = \iint \left\{\hat{f}(y\,|\,x) - f(y\,|\,x)\right\}^2 h(x)\,dx\,dy. \tag{1.5}$$

For numerical examples, we will estimate the ISE using

$$I(a,b;\mathbf{X},\mathbf{Y},\mathbf{y}',f) = \frac{\Delta}{n}\sum_{j=1}^{N}\sum_{i=1}^{n}\left[\hat{f}(y'_j\,|\,X_i) - f(y'_j\,|\,X_i)\right]^2 \tag{1.6}$$

where $\mathbf{X} = \{X_1,\ldots,X_n\}$, $\mathbf{Y} = \{Y_1,\ldots,Y_n\}$ and $\{(X_i,Y_i)\}$ is an iid sample with density $g(\cdot,\cdot)$, $\mathbf{y}' = \{y'_1,\ldots,y'_N\}$ is a vector of equally spaced values over the sample space of Y with $y_{i+1} - y_i = \Delta$, and $\hat{f}$ is calculated from $\{(X_i,Y_i)\}$ using (1.3). We average (1.6) across samples to estimate the IMSE using

$$\hat{M}(a,b;m,\mathbf{y}',f) = \frac{1}{m}\sum_{\ell=1}^{m} I(a,b;\mathbf{X}^{(\ell)},\mathbf{Y}^{(\ell)},\mathbf{y},f) \tag{1.7}$$

where $\mathbf{X}^{(\ell)} = \{X_1^{(\ell)},\ldots,X_n^{(\ell)}\}$, $\mathbf{Y}^{(\ell)} = \{Y_1^{(\ell)},\ldots,Y_n^{(\ell)}\}$, and $\{(X_i^{(\ell)},Y_i^{(\ell)})\}$ is an iid sample with density $g(\cdot,\cdot)$.

In Section 2 we derive several "reference rules" for the kernel conditional density estimator making various assumption about the conditional density $f(y\,|\,x)$ and the marginal density $h(x)$.

In Section 3 we discuss an approximate parametric bootstrap method for estimating bandwidths, similar to that used by Hall, Wolff and Yao for bandwidth selection in estimating conditional distribution functions.

A third approach is considered in Section 4, where the estimation problem is written as a regression problem so that a bandwidth selection method from kernel regression can be modified for use here.

These various approaches to bandwidth selection are combined in Section 5 to provide a practical strategy for bandwidth selection. The methods are illustrated in Section 6 using two simulated examples and one real data set. Finally, we discuss extending the bandwidth selection methods to other estimators in Section 7.

2 Reference rules

Bandwidth rules based on a reference distribution have proven useful in univariate kernel density estimation (e.g., Silverman, 1986). The most common approach is to assume the underlying density is normal and find the bandwidth which would minimize the IMSE given that assumption. This is surprisingly robust and gives reasonable results even for densities which are quite non-normal. We shall apply the idea here to obtain a quick and simple method for bandwidth selection for kernel conditional density estimators.

Hyndman et al. (1996) showed the asymptotic mean square error for the estimator $\hat{f}(y\,|\,x)$ is

$$\begin{aligned}\text{AMSE} &= \lim_{n\to\infty} \mathsf{E}\left\{\hat{f}(y\,|\,x) - f(y\,|\,x)\right\}^2 \\ &= \frac{a^4\sigma_K^4}{4}\left\{2\frac{h'(x)}{h(x)}\frac{\partial f(y\,|\,x)}{\partial x} + \frac{\partial^2 f(y\,|\,x)}{\partial x^2} + \frac{b^2}{a^2}\frac{\partial^2 f(y\,|\,x)}{\partial y^2}\right\}^2 \\ &\quad + \frac{R(K)f(y\,|\,x)}{nabh(x)}[R(K) - bf(y\,|\,x)] + O(\tfrac{1}{n}) + O(\tfrac{b}{an}) + O(\tfrac{a}{bn}) \\ &\quad + O(a^6) + O(b^6) + O(a^2b^4) + O(a^4b^2)\end{aligned} \tag{2.1}$$

where $R(K) = \int K^2(w)\,dw$.

Then substituting (2.1) into (1.4) gives

$$\text{IMSE} \approx \frac{c_1}{nab} - \frac{c_2}{na} + c_3a^4 + c_4b^4 + c_5a^2b^2 \tag{2.2}$$

where the constants c_1, c_2, c_3, c_4 and c_5 depend on the kernel K, the conditional density $f(y\,|\,x)$ and the marginal density $h(x)$. The constants are

$$\begin{aligned}c_1 &= \int R^2(K)\,dx \\ c_2 &= \iint R(K)f^2(y\,|\,x)\,dy\,dx \\ c_3 &= \iint \frac{\sigma_K^4 h(x)}{4}\left\{2\frac{h'(x)}{h(x)}\frac{\partial f(y\,|\,x)}{\partial x} + \frac{\partial^2 f(y\,|\,x)}{\partial x^2}\right\}^2 dy\,dx \\ c_4 &= \iint \frac{\sigma_K^4 h(x)}{4}\left\{\frac{\partial^2 f(y\,|\,x)}{\partial y^2}\right\}^2 dy\,dx \\ c_5 &= \iint \frac{\sigma_K^4 h(x)}{2}\left\{2\frac{h'(x)}{h(x)}\frac{\partial f(y\,|\,x)}{\partial x} + \frac{\partial^2 f(y\,|\,x)}{\partial x^2}\right\}\left\{\frac{\partial^2 f(y\,|\,x)}{\partial y^2}\right\} dy\,dx\end{aligned}$$

where the integrals are over the sample space of Y and X.

The optimal bandwidths can be derived by differentiating (2.2) with respect to a and b and setting the derivatives to zero. We require the sample space of X to be finite to ensure c_1 remains finite.

Hyndman et al. (1996) showed that the optimal bandwidths are approximately

$$a^* = c_1^{1/6}\left\{4(\frac{c_3^5}{c_4})^{1/4} + 2c_5(\frac{c_3}{c_4})^{3/4}\right\}^{-1/6} n^{-1/6} \tag{2.3}$$

$$\text{and}\quad b^* = a^*\left(\frac{c_3}{c_4}\right)^{1/4} = c_1^{1/6}\left\{4(\frac{c_4^5}{c_3})^{1/4} + 2c_5(\frac{c_4}{c_3})^{3/4}\right\}^{-1/6} n^{-1/6}. \tag{2.4}$$

We shall assume that the conditional distribution, $f(y\,|\,x)$, is normal with linear mean $r(x) = c + dx$ and linear standard deviation $\sigma(x) = p + qx$. Hence $[Y\,|\,X = x] \stackrel{d}{=} N(c + dx, (p + qx)^2)$ and the conditional density is

$$f(y\,|\,x) = \frac{1}{(p+qx)\sqrt{2\pi}}\exp\left\{\frac{-1}{2(p+qx)^2}(y - c - dx)^2\right\}.$$

We shall substitute this expression for $f(y \mid x)$ into (2.3) and (2.4) to obtain reference rules for bandwidth selection. We also need to specify the marginal density $h(x)$. We consider two possibilities: a uniform marginal density over the space $[\ell, u]$ and a normal marginal density with mean μ_h and standard deviation σ_h. The parameters of the assumed conditional and marginal distributions will be estimated for the data. When $q \neq 0$, we use iteratively reweighted least squares to estimate c and d, with p and q estimated by minimizing

$$\sum_{i=0}^{n} \left\{ (y_i - \hat{c} - \hat{d}x)^2 - (p + qx)^2 \right\}^2 .$$

To differentiate between the various assumptions about the marginal distributions and methods used we will use the following notation for the reference rules:

$$\text{bandwidth}_{[marginal][weight]}$$

where the subscripts can take the following values.

- marginal $= N$ and marginal $= U$ assume the marginal density is normal and uniform respectively.
- Weight$= i$ implies that the IMSE is weighted by $h^i(x)$. In the IMSE defined by (1.4) we have used $i = 1$. In Section 2.3 we shall consider an IMSE with $i = 2$.

For example a_{U1} denotes the optimal value of a assuming $h(x)$ is a uniform density, and the IMSE is weighted by $h(x)$. The conditional distribution is always assumed to be normal.

The derivation of each reference rule requires extensive algebraic manipulation. The following rules were obtained with some help from the computer algebra package Maple.

2.1 Uniform marginal distribution

To evaluate the reference rule for a and b with $h(x)$ uniform over $[\ell, u]$, we substitute the conditional and marginal densities into the constants $c_1, \ldots, c_5$. We then integrate the constants initially with respect to y over the sample space $(-\infty, \infty)$ and secondly with respect to x over the sample space $[\ell, u]$. The constant terms are

$$c_1 = R^2(K)(u - \ell) \qquad c_2 = \frac{R(K)}{2q\sqrt{\pi}} \log\left(\frac{p + qu}{p + q\ell}\right) \qquad c_3 = \frac{3}{512} \frac{\sigma_K^4 z w}{q\sqrt{\pi}(u - \ell)}$$

$$c_4 = \frac{3}{128} \frac{\sigma_K^4 z}{q\sqrt{\pi}(u - \ell)} \qquad c_5 = \frac{3}{128} \frac{\sigma_K^4 z(2d^2 - 3q^2)}{q\sqrt{\pi}(u - \ell)}$$

where $z = \dfrac{(p + qu)^4 - (p + q\ell)^4}{(p + qu)^4 (p + q\ell)^4}$, $w = 19q^4 + 4d^4 + 28q^2d^2$, $d \neq 0$ and $q \neq 0$.

These values are then substituted into (2.3) and (2.4) to obtain the following reference rule:

$$\begin{aligned} a_{U1} &= \left\{ \frac{2^{15/2}\sqrt{\pi}R^2(K)(u-l)^2 q}{3n\sigma_K^4 z w^{3/4}[\sqrt{w}+2d^2-3q^2]} \right\}^{1/6} \\ b_{U1} &= \frac{w^{1/4}}{\sqrt{2}} a_{U1} \end{aligned} \tag{2.5}$$

Clearly the size of the bandwidths are affected by the assumptions made on the conditional density, marginal density and the number of observations.

Suppose we now assume that the conditional standard deviation is constant (let $q = 0$). We again substitute the conditional and marginal densities into the constant terms and obtain

$$c_1 = R^2(K)(u-l) \qquad c_2 = \frac{R(K)(u-l)}{2p\sqrt{\pi}} \qquad c_3 = \frac{3}{32}\frac{\sigma_K^4 d^4}{p^5\sqrt{\pi}}$$

$$c_4 = \frac{3}{32}\frac{\sigma_K^4}{p^5\sqrt{\pi}} \qquad c_5 = \frac{3}{16}\frac{\sigma_K^4 d^2}{p^5\sqrt{\pi}}.$$

This gives the following special case of the reference rule for $q = 0$:

$$\begin{aligned} a_{U1} &= \left\{ \frac{4\sqrt{\pi}}{3}\frac{R^2(K)(u-l)p^5}{n\sigma_K^4 d^5} \right\}^{1/6} \\ b_{U1} &= d a_{U1} \end{aligned}$$

where $d \neq 0$. If $q = 0$ and $d = 0$ the conditional densities are equal for all x, and there is no need to condition on X.

2.2 Normal marginal distribution

We now assume that the marginal density $h(x)$ is normal with a constant mean μ_h and constant variance σ_h^2. We further assume that conditional density $f(y \mid x)$ has a constant variance, that is $q = 0$. (We have not been able to solve the equations for the more general case of $q \neq 0$.) Following the same procedure as for the normal-uniform reference rule we initially integrate the constants over the sample space of y $(-\infty, \infty)$. Integrating the constants over the sample space of x $(-\infty, \infty)$ for a normal marginal density results in infinite bandwidths. Therefore we choose limits of the integral over x to be $\mu_h \pm k\sigma_h$. Then we obtain

$$c_1 = 2R^2(K)k\sigma_h \qquad c_2 = \frac{R(K)k\sigma_h}{p\sqrt{\pi}} \qquad c_3 = \frac{1}{64}\frac{\sigma_K^4 d^2 v(k)}{\pi^{3/2}\sigma_h^3 p^5}$$

$$c_4 = \frac{3}{32}\frac{\operatorname{erf}\left(k/\sqrt{2}\right)\sigma_K^4 d^2}{p^5\sqrt{\pi}} \qquad c_5 = \frac{3}{16}\frac{\operatorname{erf}\left(k/\sqrt{2}\right)\sigma_K^4 d^2}{p^5\sqrt{\pi}}$$

$$\text{where} \quad v(k) = 3\operatorname{erf}\left(k/\sqrt{2}\right) d^2 {\sigma_h}^3 \pi - 8\sqrt{2\pi}\, p^2 k \exp\left(-k^2/2\right) + 8\pi p^2 \operatorname{erf}(k/\sqrt{2})$$

$$\text{and} \qquad \operatorname{erf}(x) = \tfrac{2}{\sqrt{\pi}} \int_0^x \exp(t^2/2)\, dt.$$

Substituting the constants into (2.3) and (2.4) we obtain the following reference rule:

$$\begin{aligned} a_{N1} &= \frac{\left(\dfrac{16R^2(K)k\pi^{5/4}p^5{\sigma_h}^{5/2}}{nd^{5/2}\sigma_K^4}\right)^{1/6}}{\left[\left(\dfrac{v^5(k)}{3\pi^2{\sigma_h}^4\mathrm{erf}(k/\sqrt{2})}\right)^{1/4} + 3d\left(\dfrac{v(k)\mathrm{erf}^{1/3}(k/\sqrt{2})}{3}\right)^{3/4}\right]^{1/6}} \\ b_{N1} &= \left\{\frac{d^2v(k)}{3\pi\sigma_h\mathrm{erf}(k/\sqrt{2})}\right\}^{1/4} a_{N1}. \end{aligned} \tag{2.6}$$

Numerical values of the $\mathrm{erf}(x)$ function can be computed using an approximation given in Abramowitz & Stegun (1970).

The value of k controls the size of the sample space in the x direction. Therefore as we increase k we also increase a_{N1} and b_{N1}. Common choices for k would be 2 or 3, and this would represent approximately 95% and 99.7% of the sample space respectively.

2.3 Modified IMSE function

The constant

$$c_1 = \int R^2(K)\,dx$$

will be infinite unless the sample space of X is finite. Consider the alternative weighted IMSE function

$$\mathrm{IMSE2}\{\hat{f}\} = \iint \mathsf{E}\left\{\hat{f}(y\,|\,x) - f(y\,|\,x)\right\} h^2(x)\,dx\,dy. \tag{2.7}$$

where we weight the MSE by the square of the marginal density. Then the IMSE has the same form (2.2) but with different constants. The constant c_1 becomes

$$c_1 = \int R^2(K)h(x)\,dx = R^2(K)$$

and we observe now that we no longer need to restrict the sample space x to be finite. The other constants are

$$\begin{aligned} c_2 &= \frac{R(K)}{2p\sqrt{\pi}} & c_3 &= \frac{\sigma_K^4 d^2(4p^4 + 3\sigma_h^2 d^2)}{64\pi\sigma_h^3 p^5} \\ c_4 &= \frac{3}{64}\frac{\sigma_K^4}{\sigma_h p^5\pi} & c_5 &= \frac{3}{32}\frac{\sigma_K^4 d^2}{\sigma_h p^5\pi}. \end{aligned}$$

Assuming a normal marginal density $h(x)$ with mean μ_h and variance σ_h^2, and constant conditional variance (i.e., $q = 0$), we obtain the following reference rule:

$$\begin{aligned} a_{N2} &= \left\{\frac{16\pi R^2(K){\sigma_h}^{5/2}p^5}{n\sigma_K^4 d^{5/2}\left\{(u^5/(3\sigma_h^4))^{1/4} + (3d^4u^3)^{1/4}\right\}}\right\}^{1/6} \\ b_{N2} &= \left(d^2u/(3\sigma_h^2)\right)^{1/4} a_{N2} \\ \text{where} \quad u &= (3d^2{\sigma_h}^2 + 4p^2). \end{aligned} \tag{2.8}$$

3 A bootstrap bandwidth selection approach

Following the approach of Hall, Wolff & Yao (1997) for estimation of conditional distribution functions, we propose an approximate parametric bootstrap method. We fit a parametric model

$$Y_i = \beta_0 + \beta_1 X_i + \cdots + \beta_k X_i^k + \sigma\varepsilon_i$$

where ε_i is standard normal, $\beta_0, \ldots, \beta_k$ and σ are estimated from the data and k is determined by AIC. We form a parametric estimator $\tilde{f}(y \mid x)$ based on the model. Then we simulate a bootstrap data set $\mathbf{Y}^{(\ell)} = \{Y_1^{(\ell)}, \ldots, Y_n^{(\ell)}\}$ based on the observations $\mathbf{X} = \{X_1, \ldots, X_n\}$. We choose a and b to minimize

$$\tilde{M}(a, b; m, \mathbf{y}', \tilde{f}) == \frac{1}{m}\sum_{\ell=1}^{m} I(a, b; \mathbf{X}, \mathbf{Y}^{(\ell)}, \mathbf{y}', \tilde{f}),$$

the bootstrap estimator of the IMSE (assuming the above parametric model).

This scheme is easily modified to other parametric models. For example, to allow for heteroscedasticity, replace σ by $(\sigma + vX_i)$ in the model.

4 A regression-based bandwidth selector

Fan et al. (1996) noted that the conditional density estimator $\hat{f}(y \mid x)$ obtained from (1.3) is the value of β which minimizes the weighted least squares function

$$\sum_{i=1}^{n} w_i(x)\{v_i(y) - \beta\}^2 \tag{4.1}$$

where $v_i(y) = b^{-1}K(|Y_i - y|/b)$ is a kernel function. For a given bandwidth b and a given value y, finding $\hat{f}(y \mid x)$ is a standard nonparametric problem of regressing $v_i(y)$ on X_i. Fan, Yao and Tong use this idea to define local polynomial estimators of conditional densities. We shall exploit the idea by modifying a bandwidth selection method used in regression to derive an alternative method for selecting the bandwidth a given the bandwidth b. Härdle (1991) describes selecting the bandwidth for regression by minimizing the penalized average square prediction error.

For conditional density estimation, define the penalized average square prediction error as

$$Q_b(a, y) = n^{-1}\sum_{i=1}^{n}\left\{v_i(y) - \sum_{j=1}^{n} w_j(X_i)v_j(y)\right\}^2 p(w_i(X_i)). \tag{4.2}$$

where $p(u)$ is a penalty function with first order Taylor expansion $p(u) = 1+2u+O(u^2)$. In the numerical examples in Section 6, we use Akaike's (1974) finite prediction error $p(u) = (1+u)/(1-u)$.

Substituting $\varepsilon_i = v_i(y) - f(y \mid X_i)$ into (4.2) and expanding the penalty term $p(w_i(X_i))$, we find

$$Q_b(a, y) \approx n^{-1}\sum_{i=1}^{n}\left\{\varepsilon_i + f(y \mid X_i) - \hat{f}(y \mid X_i)\right\}^2 [1 + 2w_i(X_i)]. \tag{4.3}$$

Expanding $Q_b(a,y)$ to find the leading terms and ignoring the lower order terms we obtain

$$\begin{aligned} Q_b(a,y) &= n^{-1}\sum_{i=1}^{n}\varepsilon_i^2 + ASE(a,y) + 2n^{-1}\sum_{i=1}^{n}\varepsilon_i[f(y\,|\,X_i) - \hat{f}(y\,|\,X_i)] \\ &+ 2n^{-1}\sum_{i=1}^{n}\varepsilon_i^2 w_i(X_i) + O(a^{-2}n^{-2}) + O(a^3n^{-1}) \end{aligned} \tag{4.4}$$

$$\text{where} \quad ASE(a,y) = n^{-1}\sum_{i=1}^{n}\left\{f(y\,|\,X_i) - \hat{f}(y\,|\,X_i)\right\}^2 \tag{4.5}$$

denotes the average squared error.

We now show that the third and fourth terms on the right hand side of (4.4) cancel each other out.

First we compute the conditional expectation of the third summand of $Q_b(a,y)$:

$$\begin{aligned} &\mathsf{E}\left[2n^{-1}\sum_{i=1}^{n}\varepsilon_i[f(y\,|\,X_i) - \hat{f}(y\,|\,X_i)]\,|\,X_1\ldots X_n\right] \\ &\quad= 2n^{-1}\sum_{i=1}^{n}\mathsf{E}\left[\varepsilon_i[f(y\,|\,X_i) - \sum_{j=1}^{n}w_j(X_i)v_j(y)]\,|\,X_1\ldots X_n\right] \\ &\quad= 2n^{-1}\sum_{i=1}^{n}\mathsf{E}\left[\varepsilon_i[f(y\,|\,X_i) - \sum_{j=1}^{n}w_j(X_i)(\varepsilon_j + f(y\,|\,x_j))]\,|\,X_1\ldots X_n\right] \\ &\quad= 2n^{-1}\sum_{i=1}^{n}\mathsf{E}\left[\varepsilon_i[f(y\,|\,X_i) - \sum_{j=1}^{n}w_j(X_i)f(y\,|\,x_j)]\,|\,X_1\ldots X_n\right] \\ &\qquad -2n^{-1}\sum_{i=1}^{n}\mathsf{E}\left[\varepsilon_i\sum_{j=1}^{n}w_j(X_i)\varepsilon_j\,|\,X_1\ldots X_n\right] \\ &\quad= 2n^{-1}\sum_{i=1}^{n}\mathsf{E}\left[\varepsilon_i\,|\,X_1\ldots X_n\right]\left[f(y\,|\,X_i) - \sum_{j=1}^{n}w_j(X_i)f(y\,|\,x_j)\right] \\ &\qquad -2n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}w_j(X_i)\mathsf{E}\left[\varepsilon_i\varepsilon_j\,|\,X_1\ldots X_n\right]. \end{aligned}$$

Now, $\{\varepsilon_i\}$ are independent random variables with $\mathsf{E}(\varepsilon_i\,|\,X_1,\ldots,X_n) = O(b^2)$ and variance $\sigma^2(X_i)$. The conditional expectation becomes

$$\begin{aligned} &\mathsf{E}\left[2n^{-1}\sum_{i=1}^{n}\varepsilon_i[f(y\,|\,X_i) - \hat{f}(y\,|\,X_i)]\,|\,X_1\ldots X_n\right] \\ &\qquad = -2n^{-1}\sum_{i=1}^{n}w_i(X_i)\sigma^2(X_i) + O(b^2). \end{aligned}$$

The conditional expectation of the fourth summand in (4.4) is

$$\mathsf{E}\left[2n^{-1}\sum_{i=1}^{n}\varepsilon_i^2 w_i(X_i)\,|\,X_1\ldots X_n\right] = 2n^{-1}\sum_{i=1}^{n}w_i(X_i)\sigma^2(X_i).$$

Thus, the conditional expectation of the third summand is approximately equal to the negative of the conditional expectation of the fourth summand of $Q_b(a, y)$, so that

$$Q_b(a, y) = n^{-1} \sum_{i=1}^{n} \varepsilon_i^2 + ASE(a, y) + O(b^2) + O(a^{-2}n^{-2}) + O(a^3 n^{-1}).$$

We define the penalized average squared prediction error $Q_b(a)$ as

$$\begin{aligned} Q_b(a) &= \frac{\Delta}{n} \sum_{k=1}^{N} \sum_{i=1}^{n} \left\{ \sum_{j=1}^{n} w_j(X_i)[v_i(y'_k) - v_j(y'_k)] \right\}^2 p(w_i(X_i)) \\ &= \Delta \sum_{i=1}^{n} \varepsilon_i^2 + \frac{\Delta}{N} \sum_{k=1}^{N} ASE(a, y'_k) + O(b^2) + O(a^{-2}n^{-2}) + O(a^3 n^{-1}) \end{aligned}$$

where where $\{y'_1, \ldots, y'_N\}$ are equally spaced over the sample space of Y with $y_{i+1} - y_i = \Delta$. Note that the first term in this expression is independent of a and that the second term is asymptotically equal to the IMSE defined by (1.4). Therefore, for fixed b, minimizing $Q_b(a)$ is asymptotically equivalent to minimizing the IMSE.

For computational purposes, it is convenient to write $Q_b(a)$ as

$$Q_b(a) = \frac{\Delta}{n} \mathbf{p}^T (V - W^T V) \odot (V - W^T V) \mathbf{1}$$

where V is an $n \times N$ matrix with (i, j)th element $v_i(y'_j)$, W is an $n \times n$ matrix with (i, j)th element $w_i(x_j)$, $\odot$ denotes the element-wise or Hadamard product, $\mathbf{1}$ denotes a vector of ones, and $\mathbf{p}$ denotes the vector with ith element $p(w_i(x_i))$.

5 A practical bandwidth selection strategy

The preceding sections describe several bandwidth selection methods. The reference rules are fast and easily implemented but make strong assumptions about the data. The bootstrap method is less affected by the assumed distributions, but is slow to implement. The regression-based rule usually works well in finding a value for a, but it assumes b is given.

In this section, we describe an algorithm which effectively combines these methods to provide a practical bandwidth selection strategy.

1 Find an initial value for the smoothing parameter b using one of the reference rules. For most applications, we have found rule N1 works well.

2 Given this value of b, use the regression-based method to find a value for a.

3 Use the bootstrap method to revise the estimate of b by minimizing $\tilde{M}(a, b; m, \mathbf{y}', \tilde{f})$ with respect to b while holding a fixed at the value obtained in Step 2.

Rule	a	b	IMSE ($\times 10^{-6}$)
U1($q \neq 0$)	0.94	4.7	4.0
U1($q = 0$)	0.98	4.9	3.8
N1($k = 2$)	0.74	6.8	3.5
N1($k = 3$)	0.80	7.4	3.6
N2	0.86	4.8	4.1
Regress	0.93	7.5	4.0
Bootstrap	0.87	6.5	3.4
Combination	0.91	6.7	4.0
Optimal values	0.80	7.5	3.5

Table 1: *Bandwidth estimates and IMSE values for Example 1. These are all means of 50 simulated samples each consisting of 100 observations.*

Steps 2 and 3 may be repeated one or more times. We have found this algorithm provides a relatively fast and useful approach to finding good bandwidths.

To illustrate the selection methods and the strategy described above, we shall use simulation on two examples and apply the methods to some real data.

6 Applications and comparisons

We compare the various bandwidth selection methods through two simulated models and by application to some data from the Old Faithful Geyser. In all cases, we have used the Gaussian kernel, $K(u) = \phi(u) = \exp(-u^2/2)/\sqrt{2\pi}$.

Example 1

Consider the simple model $Y_i = 10 + 5X_i + \varepsilon_i$ where $\{X_i\}$ and $\{\varepsilon_i\}$ are two independent sequences of normally distributed independent random variables with $X_1 \stackrel{d}{=} N(10, 9)$ and $\varepsilon_1 \stackrel{d}{=} N(0, 100)$. In this case, the optimal bandwidths are given by (2.6) as $a_{N1} = 0.80$ and $b_{N1} = 7.5$ (where $k = 3$).

We shall compare these with the estimated bandwidths obtained from the various methods. We shall also estimate the IMSE for each bandwidth selection method using $\hat{M}(\hat{a}, \hat{b}; m, \mathbf{y}', f)$ from (1.7) with $m = 50$, the values of $\{y'_1, \ldots, y'_N\}$ chosen to cover the interval $[-10, 130]$, $N = 25$, and $f(y \mid x) = \frac{1}{10}\phi\left(\frac{y-10-5x}{10}\right)$. For the bootstrap method, $m = 25$ is used in calculating $\tilde{M}(a, b; m, \mathbf{y}', \tilde{f})$ for each a and b.

The bandwidths and estimated IMSE obtained are given in Table 1. These are the means of 50 simulated samples each consisting of n=100 observations. Boxplots of the bandwidths and ISE values are given in Figures 2–4. It is not surprising that the N1 rule performs best as it assumes the true underlying density in this case. Note that the b values for the regression method are obtained from the N1 rule. The N2 method uses a different optimality criterion, which explains why the b values are generally smaller

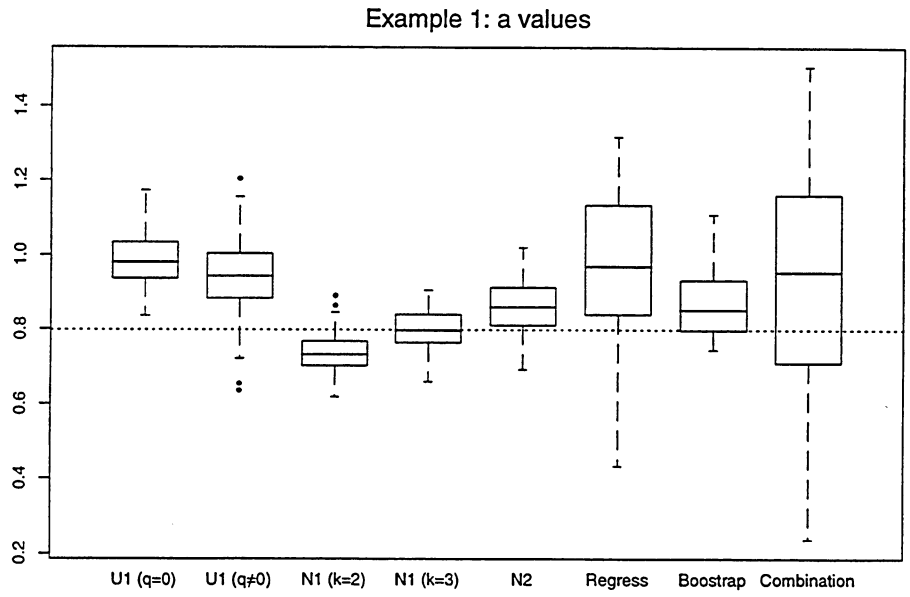

Figure 2: *Values of a for each method from 50 samples. The dotted line shows the optimal value of* $a = 0.80$.

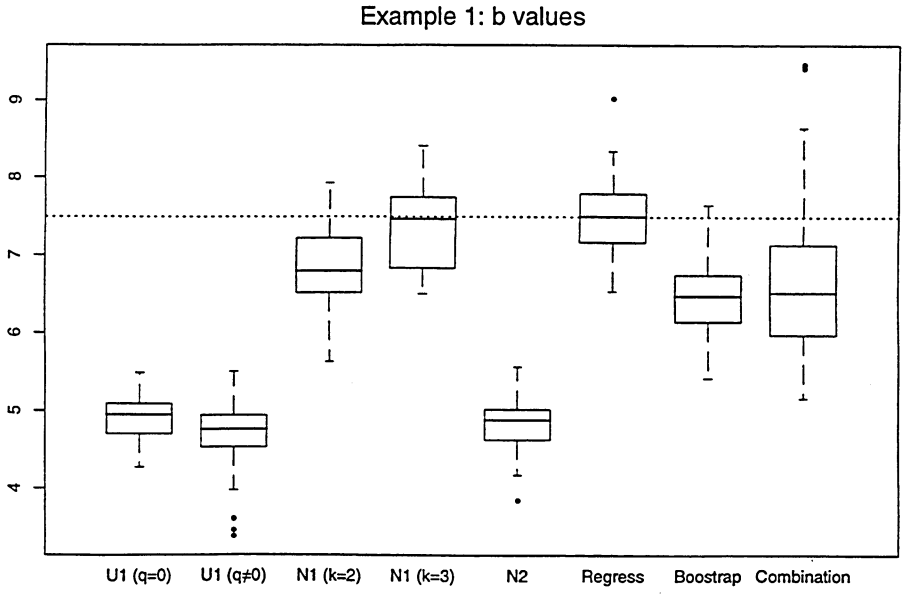

Figure 3: *Values of b for each method from 50 samples. The dotted line shows the optimal value of* $b = 7.5$.

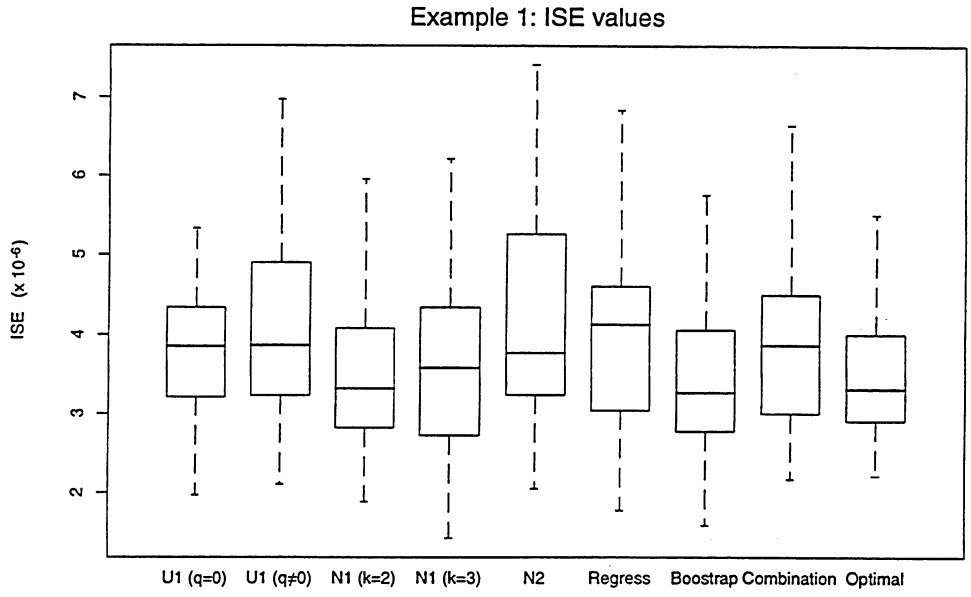

Figure 4: *Estimated ISE values for each method from 50 samples. The last boxplot shows ISE values for 50 samples using the optimal values of* $a = 0.75$ *and* $b = 7.0$.

than those obtained from the other methods. Both the bootstrap and combination methods tend to give lower values of b than the optimum. The main effect of using the combination method instead of the bootstrap method seems to be it increases the variability of the a value. However, it is much faster.

Example 2

In this example we use the model

$$Y_i = 2\sin(\pi X_i) + \varepsilon_i$$

where $\{X_i\}$ and $\{\varepsilon_i\}$ are two independent sequences of random variables with X_i uniformly distributed on $(0, 2)$ and $\varepsilon_i \,|\, X_i = W_i N_i + (1 - W_i)M_i$ where W_i is a binary variable with $\Pr(W_i = 1) = \Pr(W_i = 0) = 0.5$, $N_i \stackrel{d}{=} \mathrm{N}(X_i, 0.09)$ and $M_i \stackrel{d}{=} \mathrm{N}(0, 0.09)$. Figure 5 shows a scatterplot of 100 observations from this model.

For this model, the optimal bandwidths can be found by minimizing the estimated IMSE $\hat{M}(a, b; m, \mathbf{y}', f)$ where $f(y \,|\, x) = \frac{1}{0.6}\phi\left(\frac{y-2\sin\pi x}{0.3}\right) + \frac{1}{0.6}\phi\left(\frac{y-2\sin\pi x - x}{0.3}\right)$. Using $m = 25$, $N = 25$, and the values of $\{y'_1, \ldots, y'_N\}$ chosen to cover the interval $[-2.5, 2.5]$, we obtained optimal bandwidths of $a = 0.053$ and $b = 0.30$.

We shall also estimate the IMSE for each bandwidth selection method using (1.7) Again, we use $m = 50$ and the values of $\{y'_1, \ldots, y'_N\}$ are chosen to cover the interval $[-2.5, 2.5]$ with $N = 25$. For the bootstrap method, $m = 25$ is used.

The bandwidths and estimated IMSE obtained are given in Table 2. As for example

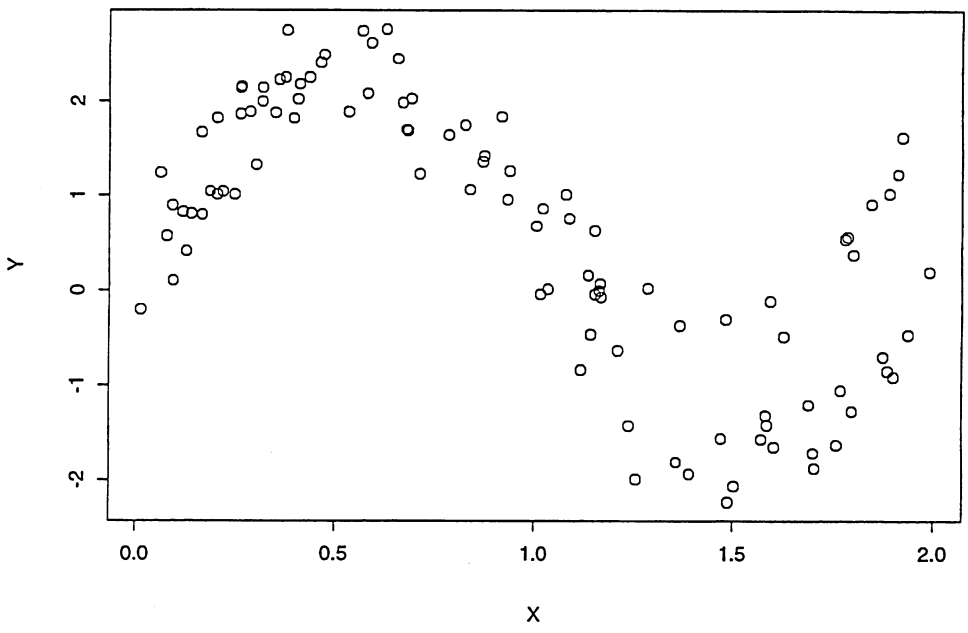

Figure 5: *Scatterplot of 100 observations from the model used in Example 2. Note the bimodality in the conditional densities for large X.*

Rule	a	b	IMSE ($\times 10^{-4}$)
U1($q \neq 0$)	0.29	0.42	3.9
U1($q = 0$)	0.32	0.45	4.1
N1($k = 2$)	0.31	0.54	4.1
N1($k = 3$)	0.31	0.58	4.2
N2	0.25	0.50	3.7
Regress	0.059	0.57	2.6
Bootstrap	0.067	0.47	2.1
Combination	0.063	0.56	2.5
Optimal values	0.053	0.30	1.8

Table 2: *Bandwidth estimates and IMSE values for Example 2. These are all means of 50 simulated samples each consisting of 100 observations.*

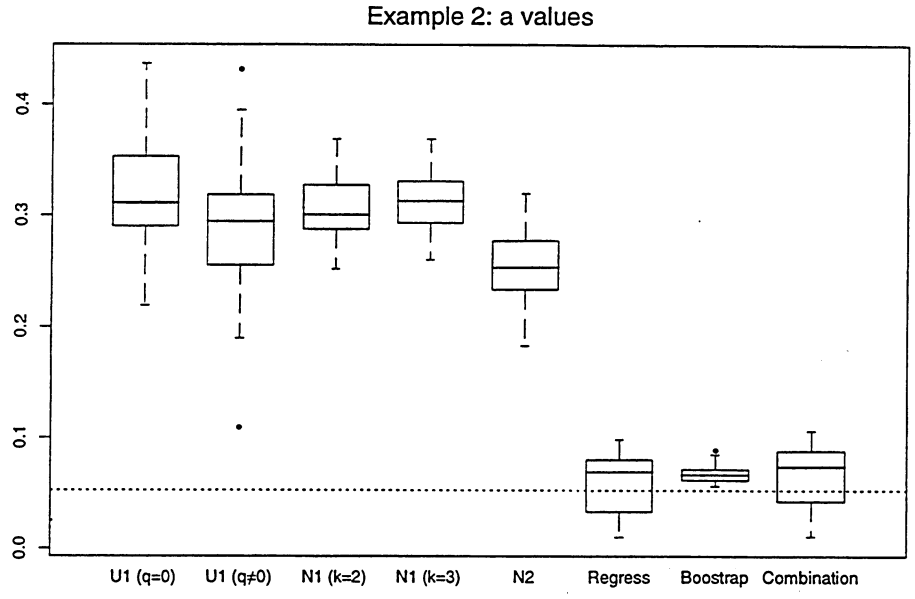

Figure 6: *Values of a for each method from 50 samples. The dotted line shows the optimal value of* $a = 0.75$.

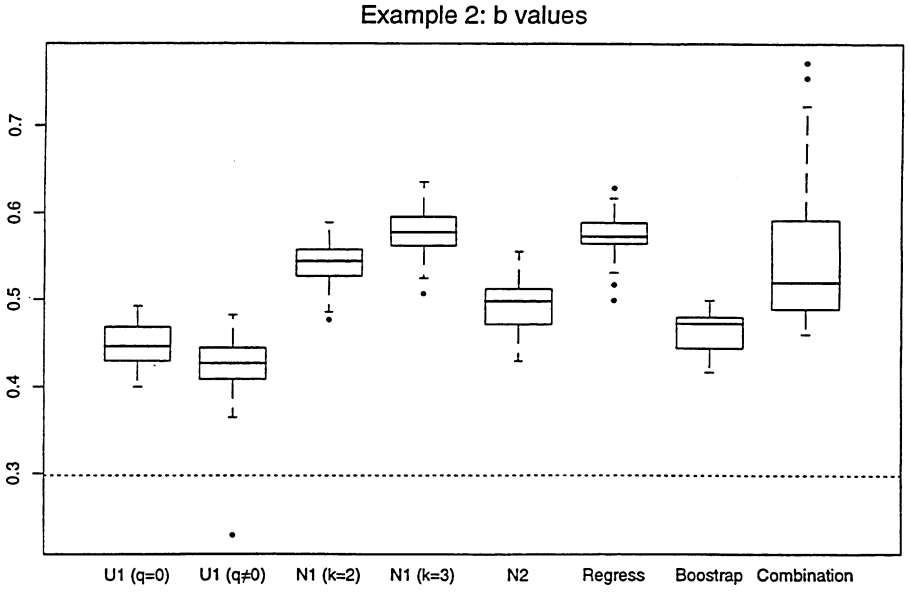

Figure 7: *Values of b for each method from 50 samples. The dotted line shows the optimal value of* $b = 7.0$.

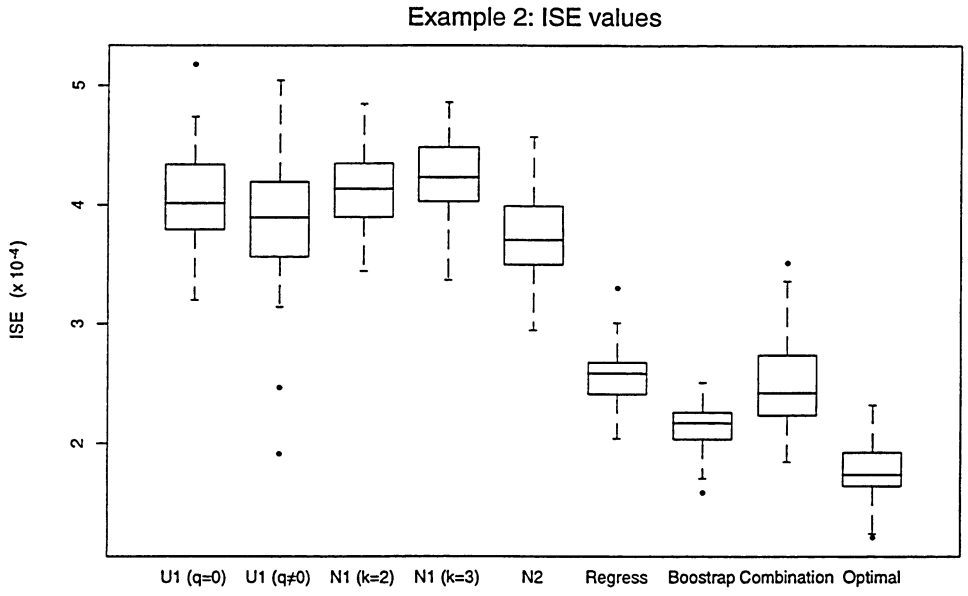

Figure 8: *Estimated ISE values for each method from 50 samples. The last boxplot shows ISE values for 50 samples using the optimal values of* $a = 0.05$ *and* $b = 0.30$.

1, these are the means of 50 simulated samples each consisting of n=100 observations. Boxplots of the bandwidths and ISE values are given in Figures 6–8. All the reference rule methods give values of a and b well above the optimum. Both the bimodality and non-linear mean of the conditional distributions lead to smaller optimal bandwidths than under the assumptions behind the reference rules. However, the regression method is still selecting values of a close to the optimum despite assuming a value of b which is much too high. The combination and bootstrap methods both lead to good values for a in this case. However, b values from both methods are too large, probably because of the assumption of normality in the bootstrap procedure. As in example 1, the combination method produces bandwidths with greater variability than the bootstrap method.

Old Faithful Geyser data

Azzalini & Bowman (1990) give data on the waiting time between the starts of successive eruptions and the duration of the subsequent eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming. The data were collected continuously from August 1st until August 15th, 1985. There are a total of 299 observations. The times are measured in minutes. Some duration measurements, taken at night, were originally recorded as S (short), M (medium), and L (long). These values have been coded as 2, 3 and 4 minutes respectively. This data set is also distributed with S-Plus.

Figure shows a scatterplot of the data. Table 3 shows the results of applying the various bandwidth selectors to these data. The conditional density estimator obtained using the bandwidth from the combination selector is shown in Figure 10.

Rule	a	b
U1($q \neq 0$)	5.1	0.27
U1($q = 0$)	6.1	0.33
N1($k = 2$)	3.9	0.81
N1($k = 3$)	4.1	0.87
N2	4.8	0.34
Regress	2.2	0.87
Bootstrap	3.6	0.40
Combination	2.4	0.48

Table 3: *Bandwidth estimates for the Old Faithful Geyser data.*

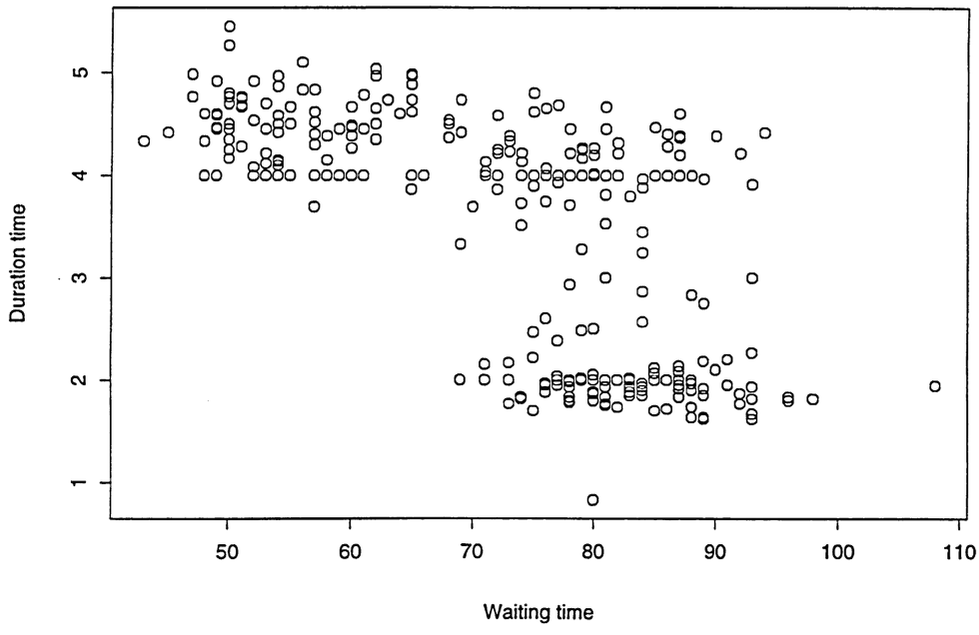

Figure 9: *Old Faithful Geyser data: duration of eruption plotted against waiting time to the eruption.*

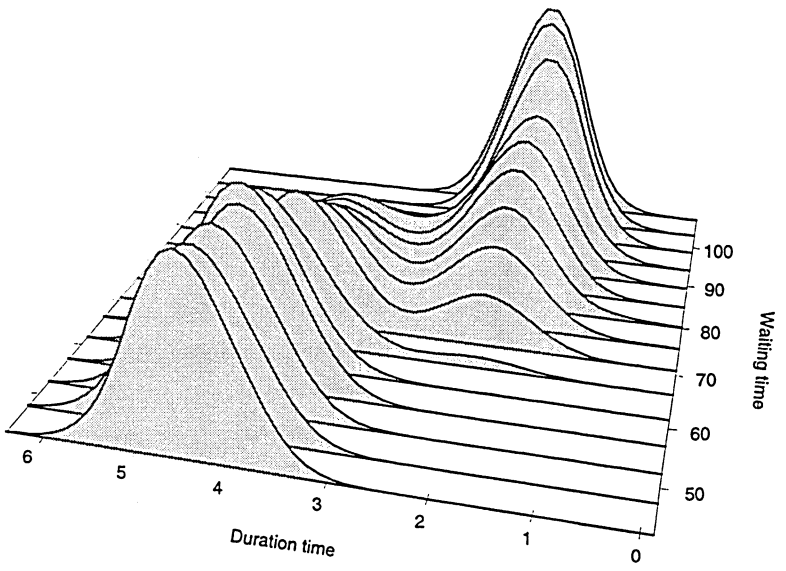

Figure 10: *Estimated conditional density of eruption duration conditional on waiting time to the eruption. Bandwidths chosen using the combination method.*

7 Extensions to other estimators

Hyndman et al. (1996) considered a modified kernel estimator which has zero mean-bias and under certain conditions a smaller IMSE. The standard kernel estimator (1.3) was modified such that the new conditional density estimator had a mean function that could be specified by a smoother with better bias properties than that inherited by the standard kernel estimator, namely the Nadaraya-Watson smoother.

Following the same approach as in Section 2, we find that the IMSE and the optimal bandwidths a^* and b^* of the modified kernel conditional density estimator take the same form as for the standard kernel conditional density estimator, except that the constants c_3 and c_5 are different:

$$c_3 = \int\int \frac{\sigma_K^4 h(x)}{4} \left\{ 2\frac{h'(x)}{h(x)} \frac{\partial f(y-r(x)|x)}{\partial x} + \frac{\partial^2 f(y-r(x)|x)}{\partial x^2} \right\}^2 dy\, dx$$

$$c_5 = \int\int \frac{\sigma_K^4 h(x)}{2} \left\{ 2\frac{h'(x)}{h(x)} \frac{\partial f(y-r(x)|x)}{\partial x} + \frac{\partial^2 f(y-r(x)|x)}{\partial x^2} \right\} \left\{ \frac{\partial^2 f(y|x)}{\partial y^2} \right\} dy\, dx,$$

where $r(x) = \mathsf{E}(Y \mid X = x)$ is the conditional mean.

Thus, reference rules can be derived for this estimator in the same way as for the standard estimator. We give just one example, the U1 rule with $q \neq 0$:

$$a_{U1} = \left\{ \frac{2^{15/2}\sqrt{\pi} R^2(K)(u-l)^2}{3(19)^{3/4} n \sigma_K^4 z q^4 (\sqrt{19}-3)} \right\}^{1/6}$$

$$\text{and} \quad b_{U1} = (19/4)^{1/4} q a_{U1}$$

where z and w are defined as for (2.5). Note that this is the same as setting $d = 0$ in (2.5).

To extend these reference rules to Fan, Yao and Tong's local polynomial estimator, one would first need to derive the IMSE of that estimator using Theorem 1 of their paper, then find expressions for the optimal values of a and b, analogous to (2.3) and (2.4). Extensions of the reference rules to the case where there is a multivariate explanatory variable is more difficult.

The bootstrap selector can be easily applied to any estimator. The regression-based selector can be adapted to other estimators (including the multivariate case) by replacing $w_j(X_i)$ by the weight from the "equivalent kernel" obtained when the estimator is written as a linear smoother.

Acknowledgments

Part of this work was carried out while Rob Hyndman was a visitor to the Department of Statistics, Colorado State University. Rob Hyndman was supported in part by an Australian Research Council grant.

References

Abramowitz, M. & Stegun, I., eds (1970), *Handbook of mathematical functions with formulas, graphs and mathematical tables*, National Bureau of Standards, Washington, D.C.

Akaike, H. (1974), 'A new look at statistical model identification', *IEEE Trans. on Automatic Control AC* **19**, 716–723.

Azzalini, A. & Bowman, A. (1990), 'A look at some data on the Old Faithful geyser', *Appl. Statist.* **39**, 357–365.

Fan, J., Yao, Q. & Tong, H. (1996), 'Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems', *Biometrika* **83**, 189–206.

Hall, P., Wolff, R. & Yao, Q. (1997), Methods for estimating a conditional distribution function, Technical report, IMS, University of Kent at Canterbury. Submitted.

Härdle, W. (1991), *Smoothing techniques with implementation in S*, Springer-Verlag, New York.

Hyndman, R., Bashtannyk, D. & Grunwald, G. (1996), 'Estimating and visualizing conditional densities', *J. Comp. Graph. Statist.*

Rosenblatt, M. (1969), Conditional probability density and regression estimators, *in* P. Krishnaiah, ed., 'Multivariate Analysis II', Academic Press, New York, pp. 25–31.

Silverman, B. (1986), *Density estimation for statistics and data analysis*, Chapman and Hall, London.

Stone, C. (1994), 'The use of polynomial splines and their tensor products in multivariate function estimation', *Ann. Statist.* **22**, 118–184.