J. R. Statist. Soc. B (1998)
60, *Part 2, pp.* 271–293

Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion

Clifford M. Hurvich and Jeffrey S. Simonoff†

New York University, USA

and Chih-Ling Tsai

University of California, Davis, USA

[Received October 1996. Revised April 1997]

Summary. Many different methods have been proposed to construct nonparametric estimates of a smooth regression function, including local polynomial, (convolution) kernel and smoothing spline estimators. Each of these estimators uses a smoothing parameter to control the amount of smoothing performed on a given data set. In this paper an improved version of a criterion based on the Akaike information criterion (AIC), termed AIC_C, is derived and examined as a way to choose the smoothing parameter. Unlike plug-in methods, AIC_C can be used to choose smoothing parameters for any linear smoother, including local quadratic and smoothing spline estimators. The use of AIC_C avoids the large variability and tendency to undersmooth (compared with the actual minimizer of average squared error) seen when other 'classical' approaches (such as generalized cross-validation or the AIC) are used to choose the smoothing parameter. Monte Carlo simulations demonstrate that the AIC_C-based smoothing parameter is competitive with a plug-in method (assuming that one exists) when the plug-in method works well but also performs well when the plug-in approach fails or is unavailable.

Keywords: Convolution kernel regression estimator; Local polynomial regression estimator; Plug-in method; Smoothing spline regression estimator

1. Introduction

Nonparametric estimation of an unknown smooth regression function has received considerable attention in recent years. Here, we shall assume that we have data $\mathbf{y} = (y_1, \ldots, y_n)'$ generated by the model

$$y_i = m(x_i) + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $m(\cdot)$ is an unknown smooth function, the x_i are given real numbers in the interval $[a, b]$ and the ϵ_i are independent random variables with mean 0 and variance σ_0^2. Either the predictor vector $\mathbf{x}$ is non-random, or analyses proceed conditionally on the observed values if it is random.

Many different estimators of m have been proposed. A pth-order local polynomial estimator is defined as the constant term $\hat{\beta}_0$ of the minimizer of

†*Address for correspondence*: Department of Statistics and Operations Research, Leonard N. Stern School of Business, New York University, Room 8-54, 44 West 4th Street, New York, NY 10012-1126, USA.
E-mail: jsimonoff@stern.nyu.edu

 1369–7412/98/60271

$$\sum_{i=1}^{n} \{y_i - \beta_0 - \ldots - \beta_p(x - x_i)^p\}^2 K\left(\frac{x - x_i}{h}\right),$$

where K is the kernel function, generally taken to be a symmetric probability density function with finite second derivative (for a general discussion of this estimator, as well as the other techniques described in this section, see Simonoff (1996), chapter 5). Typical choices of p are 0, 1, 2 and 3, with certain asymptotic and boundary bias correction advantages going to the local linear ($p = 1$) and local cubic ($p = 3$) estimators over the local constant ($p = 0$) and local quadratic ($p = 2$) estimators respectively. Higher values of p (2 or 3) also can take advantage of greater smoothness of m by yielding a faster convergence rate to 0 of the mean-squared error (MSE) of the estimator, $\text{MSE} = E[\{\hat{m}(x) - m(x)\}^2]$.

A Gasser–Müller convolution kernel estimator (Gasser and Müller, 1979) takes the form

$$\hat{m}(x) = h^{-1} \sum_{i=1}^{n} \left\{ \int_{s_{i-1}}^{s_i} K\left(\frac{x - u}{h}\right) \mathrm{d}u \right\} y_i, \tag{1.1}$$

where $x_{i-1} \leqslant s_{i-1} \leqslant x_i$ (a common choice being $s_{i-1} = (x_{i-1} + x_i)/2$, with s_0 and s_n being the upper and lower limits of the range of x respectively). Here the kernel function K need not be a probability density function, as so-called higher order kernels can yield improved MSE convergence rates for smoother m (analogously to local quadratic and cubic estimators). The kernel functions must be corrected for potential bias effects in the boundary regions of the data by using boundary kernels.

A third approach to estimating m is by using smoothing splines. A cubic smoothing spline estimator is the minimizer of

$$\frac{1}{n} \sum_{i=1}^{n} \{y_i - m(x_i)\}^2 + \alpha \int m''(u)^2 \, \mathrm{d}u$$

over the class of functions with m and m' absolutely continuous and m'' square integrable.

Although these estimators are defined in different ways, there are connections between them. For example, for a fixed design of equidistant values of $\mathbf{x}$, local polynomial and convolution kernel estimators are asymptotically equivalent in the interior (and at the boundary if boundary kernels are used). Despite this, the finite sample properties of the estimators can be very different. A property that is key to the derivations in this paper is that all the estimators are linear, in that $\hat{\mathbf{y}} = \hat{m}(\mathbf{x}) = H\mathbf{y}$, where the matrix H is commonly called the hat matrix or smoother matrix and depends on $\mathbf{x}$ but not on $\mathbf{y}$ (regression spline and wavelet estimators are also linear estimators).

A crucial step in estimating m is choosing the smoothing parameter (h for the local polynomial and kernel estimators, α for the smoothing spline), which controls the smoothness of the resultant estimate. Automatic smoothing parameter selectors generally fall into two broad classes of methods: classical and plug-in approaches. Classical methods are based on the minimization of an approximately unbiased estimator of either the mean average squared error

$$\text{MASE} = \frac{1}{n} E[(\hat{\mathbf{m}}_h - \mathbf{m})'(\hat{\mathbf{m}}_h - \mathbf{m})]$$

(e.g. generalized cross-validation (GCV); Craven and Wahba (1979)) or the expected Kullback–Leibler discrepancy given in equation (2.1) (e.g. the Akaike information criterion (AIC); Akaike (1973)). Here we use the shorthand notation $\mathbf{m}$ to represent $m(\mathbf{x}) =$

$m(x_1, \ldots, x_n)'$ and use h as a generic smoothing parameter for any linear smoother (including the smoothing spline). The smoothing parameter is chosen to be the minimizer of $\log(\hat{\sigma}^2) + \psi(H)$, where

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{m}_h(x_i)\}^2 = \mathbf{y}'(I-H)'(I-H)\mathbf{y}/n$$

and $\psi(\cdot)$ is a penalty function designed to decrease with increasing smoothness of $\hat{\mathbf{m}}_h$. Common choices of ψ lead to GCV ($\psi(H) = -2\log\{1 - \mathrm{tr}(H)/n\}$), the AIC ($\psi(H) = 2\,\mathrm{tr}(H)/n$) and T (Rice, 1984) ($\psi(H) = -\log\{1 - 2\,\mathrm{tr}(H)/n\}$). Each of these selectors depends on H through its trace, which can be interpreted as the effective number of parameters used in the smoothing fit (see, for example, Hastie and Tibshirani (1990), section 3.5).

Classical bandwidth selectors (particularly GCV and the AIC) have to some extent fallen into disuse (particularly in application to local polynomial and kernel estimators) because of two unfavourable properties: the selectors lead to highly variable choices of smoothing parameter, and they have a noticeable tendency towards undersmoothing (too large a value of $\mathrm{tr}(H)$). These difficulties have inspired the formulation of plug-in methods.

The plug-in selector of Ruppert *et al.* (1995) for the local linear estimator is typical. It can be shown that the bandwidth that minimizes the weighted conditional mean integrated squared error

$$\mathrm{MISE}(\hat{m}|x_1, \ldots, x_n) = E\left[\int_a^b \{\hat{m}(u) - m(u)\}^2 f_X(u)\,du | x_1, \ldots, x_n\right]$$

asymptotically is

$$h_{0,a} = \left[R(K)\sigma_0^2 \Big/ \left\{n\,\mu_2(K)^2 \int m''(u)^2 f_X(u)\,du\right\}\right]^{1/5}, \tag{1.2}$$

where $R(K) \equiv \int K(u)^2\,du$, $\mu_2(K) \equiv \int u^2 K(u)\,du$ and $f_X(u)$ is the density function for the predictors (results for fixed designs take $x_i = F_X^{-1}(i/n)$, with F_X the cumulative function of the 'density' f_X of the design). The plug-in bandwidth is given by the right-hand side of equation (1.2) with estimates of σ_0^2 and $\int m''(u)^2 f_X(u)\,du$ substituted for the actual values. The resultant selector is much less variable than that based on GCV and does not tend to undersmooth in practice.

Despite these favourable properties, plug-in selectors have several theoretical and practical problems. First, they only have been defined where the asymptotically optimal bandwidth $h_{0,a}$ has a simple form, which is not the case for the local quadratic estimator (for that estimator $h_{0,a}$ depends on m''', $m^{(\mathrm{iv})}$ and f_X'). Similarly, no plug-in methods have been proposed for smoothing splines.

Plug-in selectors also have philosophical drawbacks. The main theoretical advantages of plug-in selectors over classical selectors refer to estimation of h_0, the bandwidth that minimizes MISE for the given sample size and design. This bandwidth, which approaches $h_{0,a}$ as $n \to \infty$, is thus optimal with respect to the average performance over all possible data sets for a given population, rather than the performance for the observed data set. Although plug-in selectors are far better at estimating h_0 than are classical selectors, these large advantages do not carry over to estimation of $\hat{h}_0$, the bandwidth that minimizes the integrated squared error or average squared error (ASE) for the observed data set. In our opinion $\hat{h}_0$ is a more reasonable target from a conceptual point of view, and therefore many of the reported theoretical advantages of plug-in selectors do not refer to a question that is relevant to the data analyst. See Mammen (1990), Hall and Marron (1991), Jones (1991), Jones and

Kappenman (1991) and Grund *et al.* (1994) for further discussion of the issues in estimating h_0 *versus* $\hat{h}_0$.

The local linear plug-in selector proceeds by estimating $\int m''(u)^2 f_X(u)\,du$, which requires the assumption that roughly four continuous derivatives for m exist. That much smoothness renders the local linear estimator itself asymptotically inefficient, however, calling into question the entire operation. This point was noted by Terrell (1992) and Loader (1995), among others. Estimating this functional typically requires the data analyst to choose preliminary parameters in either a data-dependent or fixed fashion, and the properties of the final plug-in bandwidth can be sensitive to these choices.

In this paper classical smoothing parameter selectors based on improved versions of the AIC are proposed. As is true for all classical methods, the selectors are defined for all linear estimators. Moreover, those proposed here do not exhibit the high variability and tendency to undersmoothing of GCV (Hart and Yi (1996) proposed a variant of cross-validation with the same goal in mind). The derivations are given in Section 2. Monte Carlo results discussed in Section 3 show that one of the improved selectors, $\mathrm{AIC_C}$, performs comparably with well-behaved plug-in methods, while also performing well when the plug-in selectors fail. $\mathrm{AIC_C}$ is based on the smoother only through $\mathrm{tr}(H)$, so it is as easy to apply as GCV, the AIC and T. Section 3 also includes an application to a real data set, while Section 4 discusses possible future work.

2. Improved versions of Akaike's information criterion for smoothing parameter selection

The AIC was originally designed for parametric models as an approximately unbiased estimate of the expected Kullback–Leibler information. For linear regression and time series models, Hurvich and Tsai (1989) demonstrated that in small samples the bias of the AIC can be quite large, especially as the dimension of the candidate model approaches the sample size (thus leading to overfitting of the model), and they proposed a corrected version, $\mathrm{AIC_C}$, which was found to be less biased than the AIC. We shall now develop two criteria, $\mathrm{AIC_{C_0}}$ and $\mathrm{AIC_{C_1}}$, which are specifically designed as approximately unbiased estimates of expected Kullback–Leibler information in the context of nonparametric regression. $\mathrm{AIC_{C_0}}$ is the more exact of the two, but requires numerical integration for its evaluation. $\mathrm{AIC_{C_1}}$ is an approximation to $\mathrm{AIC_{C_0}}$ which is simpler to evaluate (although it requires calculations involving all the elements of an $n \times n$ matrix), and which is found in practice to perform identically with $\mathrm{AIC_{C_0}}$. A third criterion, $\mathrm{AIC_C}$, is an approximation to $\mathrm{AIC_{C_1}}$ that is as simple to apply as the classical criteria discussed in Section 1.

We now present the derivation of $\mathrm{AIC_{C_0}}$. Given data $\mathbf{y}$ generated from the true model $\mathbf{y} = \mathbf{m} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim N(0, \sigma_0^2 I_n)$, we consider the candidate model $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\eta}$, where $\boldsymbol{\eta} \sim N(0, \sigma^2 I_n)$. It should be stressed that in spite of the normality assumption imposed here the resulting criteria exhibit good performance in simulations, even for non-normal errors. If $f(\mathbf{y})$ denotes the likelihood for $(\boldsymbol{\mu}, \sigma^2)$ and E_0 denotes expectation with respect to the true model, we consider the Kullback–Leibler discrepancy function

$$\begin{aligned} d(\boldsymbol{\mu}, \sigma^2) &= E_0[-2 \log\{f(\mathbf{y})\}] \\ &= n \log(2\pi\sigma^2) + E_0[(\mathbf{m} + \boldsymbol{\epsilon} - \boldsymbol{\mu})'(\mathbf{m} + \boldsymbol{\epsilon} - \boldsymbol{\mu})/\sigma^2] \\ &= n \log(2\pi\sigma^2) + n\frac{\sigma_0^2}{\sigma^2} + (\mathbf{m} - \boldsymbol{\mu})'(\mathbf{m} - \boldsymbol{\mu})/\sigma^2. \end{aligned}$$

Thus,

$$d(\hat{\mathbf{m}}_h, \hat{\sigma}^2) = n \log(2\pi\hat{\sigma}^2) + n\frac{\sigma_0^2}{\hat{\sigma}^2} + (\mathbf{m} - \hat{\mathbf{m}}_h)'(\mathbf{m} - \hat{\mathbf{m}}_h)/\hat{\sigma}^2.$$

A reasonable criterion for judging the quality of the estimator $\hat{\mathbf{m}}_h$ in the light of the data is $\Delta(h) = E_0[d(\hat{\mathbf{m}}_h, \hat{\sigma}^2)]$. Ignoring the constant $n \log(2\pi)$, we have

$$\Delta(h) = E_0[n \log(\hat{\sigma}^2)] + n\sigma_0^2 E_0[1/\hat{\sigma}^2] + E_0[(\mathbf{m} - \hat{\mathbf{m}}_h)'(\mathbf{m} - \hat{\mathbf{m}}_h)/\hat{\sigma}^2]. \tag{2.1}$$

Unfortunately, $\Delta(h)$ will not be known in practice, since it depends on the true regression function $\mathbf{m}$. Therefore, we seek an approximately unbiased estimator of $\Delta(h)$ which depends only on the observed data $\mathbf{y}$. At this stage, it is helpful to make the simplifying assumption that $\hat{\mathbf{m}}_h$ is unbiased, i.e. $E_0[\hat{\mathbf{m}}_h] = \mathbf{m}$, or equivalently $H\mathbf{m} = \mathbf{m}$. Clearly, this assumption will rarely hold exactly in practice. Nevertheless, Cleveland and Devlin (1988) made a similar assumption in their derivation of a nonparametric analogue of Mallows's C_p. The assumption of unbiasedness, which is needed only to facilitate the derivation of a feasible penalty function, plays an analogous role to the key simplifying assumption used in the derivation of the AIC for parametric models, namely that the candidate family of models includes the true model (see Akaike (1974) and Linhart and Zucchini (1986), p. 245). It should be stressed that the assumption is made only in the derivation of the criterion. It is then possible to study the performance of this criterion without regard to the assumptions underlying its derivation.

Assuming, then, that $H\mathbf{m} = \mathbf{m}$, $\Delta(h)$ reduces to

$$\tilde{\Delta}(h) = E_0[n \log(\hat{\sigma}^2)] + n^2 E_0\left[\frac{\sigma_0^2}{\epsilon'(I-H)'(I-H)\epsilon}\right] + n E_0\left[\frac{\epsilon' H' H \epsilon}{\epsilon'(I-H)'(I-H)\epsilon}\right]. \tag{2.2}$$

Even if $\hat{\mathbf{m}}_h$ is not unbiased, $\tilde{\Delta}(h)$ serves as an approximation to $\Delta(h)$. Let $B_1 = (I-H)'(I-H)$, and write $B_1 = \Gamma D \Gamma'$, where D is a diagonal matrix of eigenvalues of B_1 and Γ is an orthogonal matrix whose columns are the corresponding eigenvectors of B_1. Also, let $B_2 = H'H$ and $C = \Gamma' B_2 \Gamma$. Define $\mathbf{z} = \Gamma'\epsilon/\sigma_0$. Let A_1 be the second term on the right-hand side of equation (2.2) and A_2 be the final term. Using results in Jones (1986, 1987), it can be shown that

$$A_1 = n^2 \int_0^1 (1-t)^{r/2-2} \prod_{j=1}^{r} (1 - t + 2td_j)^{-1/2}\, \mathrm{d}t, \tag{2.3}$$

where r is the rank of B_1 and d_j is the jth diagonal element of D, and

$$A_2 = n \int_0^\infty \sum_{i=1}^{n} \frac{c_{ii}}{1 + 2d_i t} \prod_{i=1}^{n} (1 + 2d_i t)^{-1/2}\, \mathrm{d}t, \tag{2.4}$$

where c_{ii} is the ith diagonal element of C.

AIC_{C_0} is defined as

$$\mathrm{AIC}_{C_0} = n \log(\hat{\sigma}^2) + A_1 + A_2,$$

where A_1 and A_2 may be obtained from formulae (2.3) and (2.4) by (one-dimensional) numerical integration. AIC_{C_0} is exactly unbiased for $\tilde{\Delta}(h)$, regardless of whether $H\mathbf{m} = \mathbf{m}$ holds, but if $H\mathbf{m} \neq \mathbf{m}$ then $\tilde{\Delta}(h)$ will not coincide exactly with the true expected Kullback–Leibler information $\Delta(h)$. The situation here is quite analogous to that for model selection in linear

regression, where the AIC is typically biased (even asymptotically) when the dimension of the candidate model is less than the dimension of the true model.

Although the terms A_1 and A_2 in AIC_{C_0} are easily and accurately obtained, the necessity for using numerical integration (even in one dimension), as well as numerical eigensystem routines, may be considered a drawback. We therefore present another criterion, AIC_{C_1}, which can be evaluated without resort to numerical integration, and which is found to provide an excellent approximation to AIC_{C_0}. From equation (2.2) and the notation following it, we may write

$$\tilde{\Delta}(h) = E_0[n \log(\hat{\sigma}^2)] + n^2 E_0[1/\mathbf{z}'D\mathbf{z}] + n E_0[\mathbf{z}'C\mathbf{z}/\mathbf{z}'D\mathbf{z}].$$

Using the method described by Cleveland and Devlin (1988) based on Satterthwaite's approximation (Khatri (1980) and Kotz and Johnson (1986), pages 376–379), the distributions of $\mathbf{z}'D\mathbf{z}$ and $\mathbf{z}'C\mathbf{z}/\mathbf{z}'D\mathbf{z}$ are approximated as

$$\mathbf{z}'D\mathbf{z} \sim (\delta_2/\delta_1)\chi^2_{\delta_1^2/\delta_2}$$

and

$$\mathbf{z}'C\mathbf{z}/\mathbf{z}'D\mathbf{z} \sim (\nu_1/\delta_1)F_{\nu_1^2/\nu_2,\delta_1^2/\delta_2},$$

where $\delta_1 = \mathrm{tr}(B_1)$, $\delta_2 = \mathrm{tr}(B_1^2)$, $\nu_1 = \mathrm{tr}(B_2)$, $\nu_2 = \mathrm{tr}(B_2^2)$ and B_1 and B_2 are as defined above. Treating these distributional approximations as exact yields

$$E_0[1/\mathbf{z}'D\mathbf{z}] = \frac{\delta_1/\delta_2}{\delta_1^2/\delta_2 - 2},$$

and

$$E_0[\mathbf{z}'C\mathbf{z}/\mathbf{z}'D\mathbf{z}] = \frac{\nu_1\delta_1/\delta_2}{\delta_1^2/\delta_2 - 2}.$$

AIC_{C_1}, proposed as an approximately unbiased estimator of $\tilde{\Delta}(h)$, is defined as

$$\mathrm{AIC}_{C_1} = n \log(\hat{\sigma}^2) + n \frac{(\delta_1/\delta_2)(n + \nu_1)}{\delta_1^2/\delta_2 - 2}.$$

The accuracy of the approximation of AIC_{C_1} to AIC_{C_0} was examined in Monte Carlo simulations (not reported here) and was found to be excellent.

AIC_{C_0} and AIC_{C_1} are somewhat complicated to apply in practice, as they require eigenanalysis and numerical integration in the former case and calculations involving all the elements of the $n \times n$ matrix H in the latter case (these calculations can be accelerated by using binning techniques; see Turlach and Wand (1996)). Hurvich and Tsai (1989) showed that in parametric linear regression and autoregressive time series contexts the bias-corrected AIC (AIC_C) takes the form

$$\log(\hat{\sigma}^2) + \frac{1 + p/n}{1 - (p+2)/n} = \log(\hat{\sigma}^2) + 1 + \frac{2(p+1)}{n - p - 2},$$

where $\hat{\sigma}^2$ is the estimated error (or innovations) variance and p is the number of regression (or autoregressive) parameters in the model. By analogy, then, we obtain the version of AIC_C for smoothing parameter selection,

$$\mathrm{AIC_C} = \log(\hat{\sigma}^2) + \frac{1 + \mathrm{tr}(H)/n}{1 - \{\mathrm{tr}(H) + 2\}/n} = \log(\hat{\sigma}^2) + 1 + \frac{2\{\mathrm{tr}(H) + 1\}}{n - \mathrm{tr}(H) - 2}. \tag{2.5}$$

This criterion is easier to apply, as it is a function of H only through its trace.

If H is assumed to be symmetric and idempotent (an assumption which was not made in the derivation of $\mathrm{AIC_{C_1}}$), then $\mathrm{AIC_{C_1}}$ reduces to $\mathrm{AIC_C}$. Since H will not be symmetric and idempotent in general (although it is for regression splines), one way to think of $\mathrm{AIC_C}$ is as an approximation to $\mathrm{AIC_{C_1}}$ (which is, in turn, a very accurate approximation to $\mathrm{AIC_{C_0}}$). Negative penalties in $\mathrm{AIC_{C_1}}$ and $\mathrm{AIC_C}$ are treated as infinite.

It follows from Härdle *et al.* (1988) that all the classical selectors considered here are asymptotically equivalent. Given this, we might wonder why they might exhibit noticeably different performances in practice. The reason is that the asymptotic theory assumes that $\mathrm{tr}(H)/n \to 0$, a situation that is not consistent with a small smoothing parameter.

Fig. 1 makes this distinction clear. It gives the penalty functions $\psi(H)$ as a function of $\mathrm{tr}(H)$ for GCV, T, the AIC and $\mathrm{AIC_C} - 1$ (subtracting 1 from $\mathrm{AIC_C}$ makes it comparable with the other selectors, as can be seen from equation (2.5), and does not affect its smoothing parameter choices; since $\mathrm{AIC_C}$ depends on n, its curve is given for $n = 100$). All four ψ-functions become indistinguishable at the left-hand end of the plot, which corresponds to $\mathrm{tr}(H)/n \to 0$ and the usual asymptotics. The criteria differ markedly for a small smoothing parameter (large $\mathrm{tr}(H)/n$), however, with a sharper rise corresponding to a heavier penalty against undersmoothing. The AIC and GCV have relatively weak penalties; this accounts for

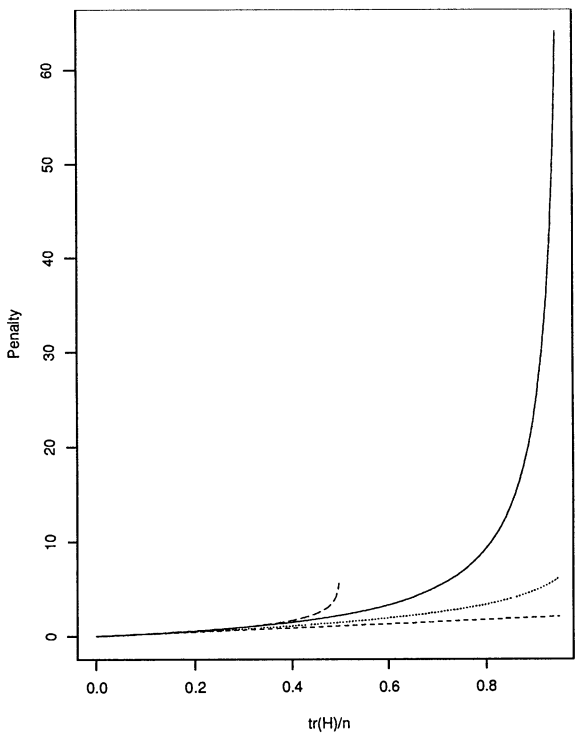


Fig. 1. ψ-penalties for various selectors as a function of $\mathrm{tr}(H)/n$: ———, $\mathrm{AIC_C}$; ·····, GCV; – – –, T; - - - -, AIC

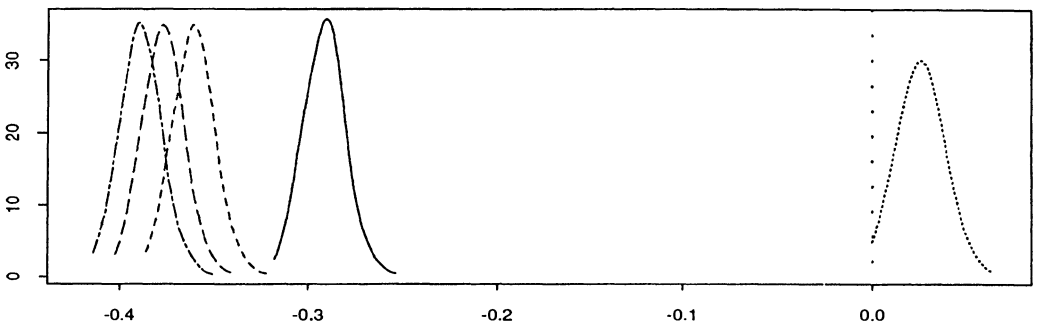
their tendencies to lead to undersmoothing. T, in contrast, has a very strong penalty, as it is effectively infinite for $\mathrm{tr}(H)/n \geqslant 0.5$. This means that T must lead to oversmoothing when a very small smoothing parameter is appropriate. $\mathrm{AIC_C}$ occupies a position between these two extremes, being less susceptible to both the undersmoothing of the AIC and GCV and the oversmoothing of T.

3. Practical performance of the selectors

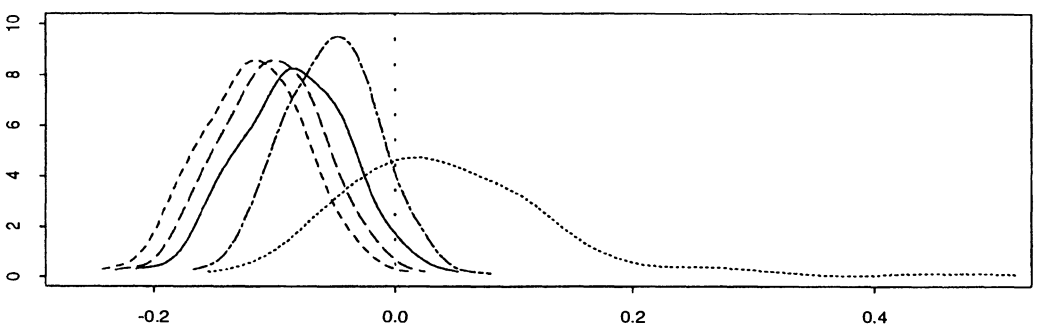
In this section we use Monte Carlo simulations and real data examples to investigate the properties of the various selectors in practice. The Monte Carlo simulations examine the performance of the selectors as they relate to the sample size, the pattern of predictor values, the true regression function, the true standard deviation of the errors and the regression estimator being used. Although only some of the results are reported here, the following settings of these factors were examined, with 500 simulation replications for each setting of factors:

(a) sample size $n = 50$, 100 and 500;
(b) the pattern of predictor values — an equispaced fixed design, a random uniform design and a non-uniform fixed design, all on [0, 1];
(c) the following six regression functions, most of which were used in earlier Monte Carlo studies (Ruppert *et al.*, 1995; Hart and Yi, 1996; Herrmann, 1997) —
  (i) $m(x) = \sin(15\pi x)$ (a function with a large amount of fine structure),
  (ii) $m(x) = \sin(5\pi x)$ (a function with less fine structure),
  (iii) $m(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$ (a function with less fine structure and a trend, in some sense 'typical' of many regression situations),
  (iv) $m(x) = 0.3\ \exp\{-64(x - 0.25)^2\} + 0.7\ \exp\{-256(x - 0.75)^2\}$ (a function with noticeably different degrees of curvature for different values of the predictor),
  (v) $m(x) = 10\ \exp(-10x)$ (a function with a trend, but no fine structure),
  (vi) $m(x) = \exp(x - \frac{1}{3})$ for $x < \frac{1}{3}$ and $\exp\{-2(x - \frac{1}{3})\}$ for $x \geqslant \frac{1}{3}$ (a function with undefined first derivative at $x = \frac{1}{3}$, which violates the standard assumptions for optimal performance of the estimators used here);
(d) error standard deviation $\sigma_0 = 0.01R_y$, $0.05R_y$, $0.25R_y$ and $0.5R_y$, where R_y is the range of $m(x)$ over $x \in [0, 1]$;
(e) regression estimators — the local linear and quadratic estimators using a Gaussian kernel, second-order and fourth-order boundary-corrected Gasser–Müller convolution kernel estimators, as described in Herrmann (1997), and a cubic smoothing spline estimator.

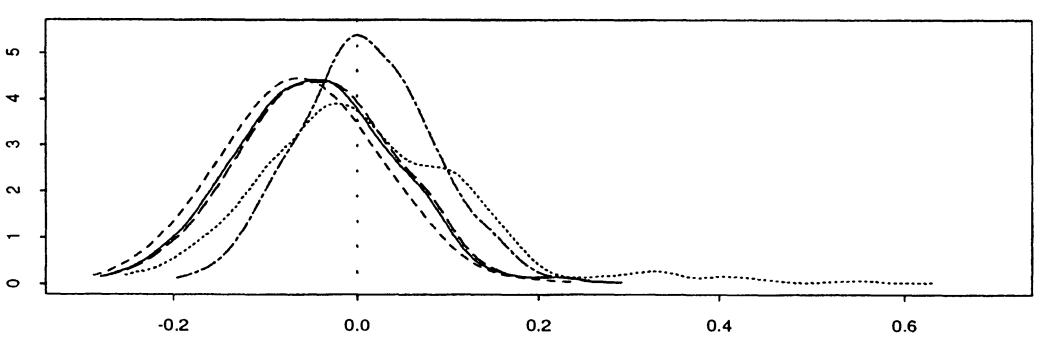
Tables 1–4 and Figs 2–6 summarize some of the results of the simulations. Tables 1–4 give the average of the optimal ASEs (based on the optimal smoothing parameter for the given simulated data set) and the mean of the ratio of the ASE to the optimal value when using a particular selector. Squared error does not completely reflect the actual performance of the selectors, so the distributions of the amount of smoothing done by the selectors compared with the smoothing done by the smoothing parameters that minimize the ASE are also examined in several figures. Selectors based on GCV, T, $\mathrm{AIC_C}$ and $\mathrm{AIC_{C_1}}$ are reported for all estimators. In addition, results for the plug-in selector described in Herrmann (1997) for each of the convolution kernel estimators are given, as are results for the local linear plug-in selector of Ruppert *et al.* (1995). Finally, a plug-in selector for the local quadratic estimator is calculated as $27\hat{h}_L/16$, where $\hat{h}_L$ is the local linear plug-in bandwidth, since that yields a



(a)

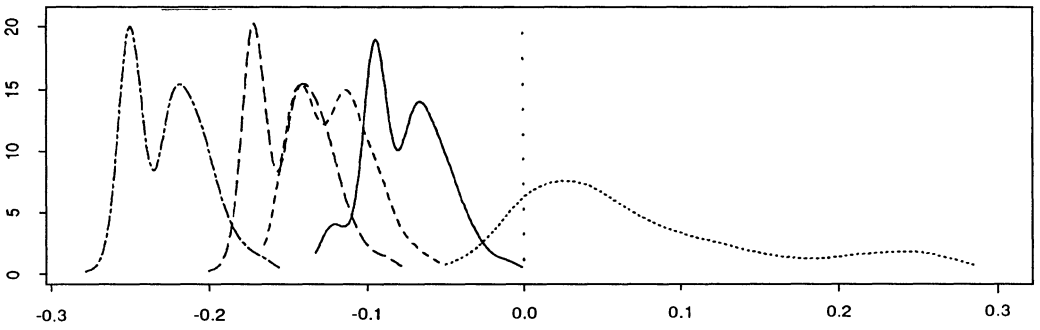


(b)

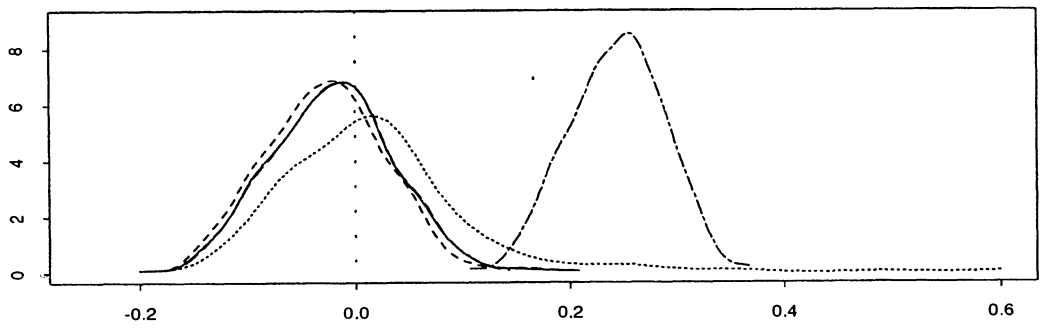


(c)

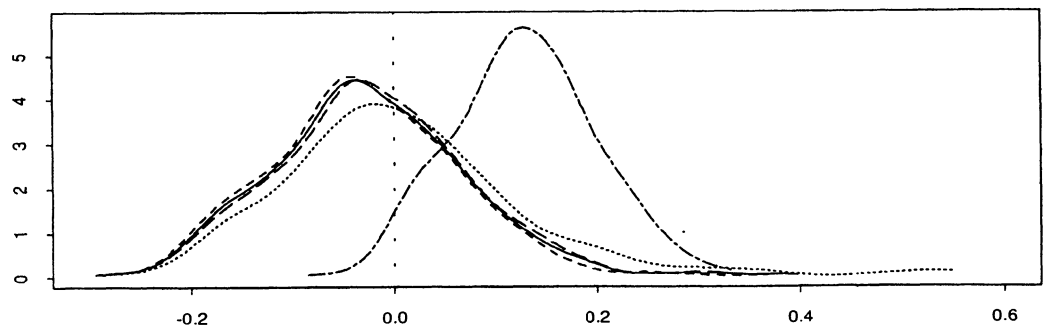
Fig. 2. Kernel density estimates of the distribution of logged ratios of the number of fitted parameters to the number corresponding to the minimizer of the ASE for the local linear estimator using AIC_C (———), GCV (·····), T (– – –), AIC_{C_1} (- - - -) and the plug-in method (—·—): (a) $m(x) = \sin(15\pi x)$, $\sigma/R_y = 0.01$; (b) $m(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$, $\sigma/R_y = 0.01$; (c) $m(x) = 10\exp(-10x)$, $\sigma/R_y = 0.05$ (negative values indicate oversmoothing; positive values indicate undersmoothing)



(a)

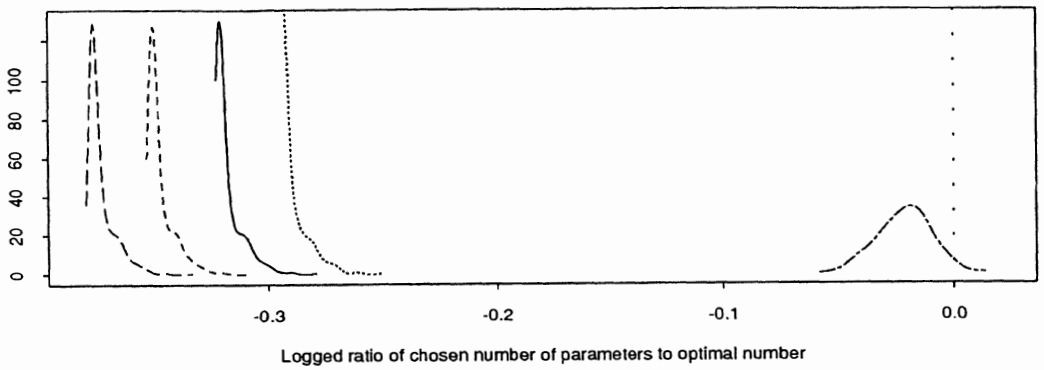


(b)

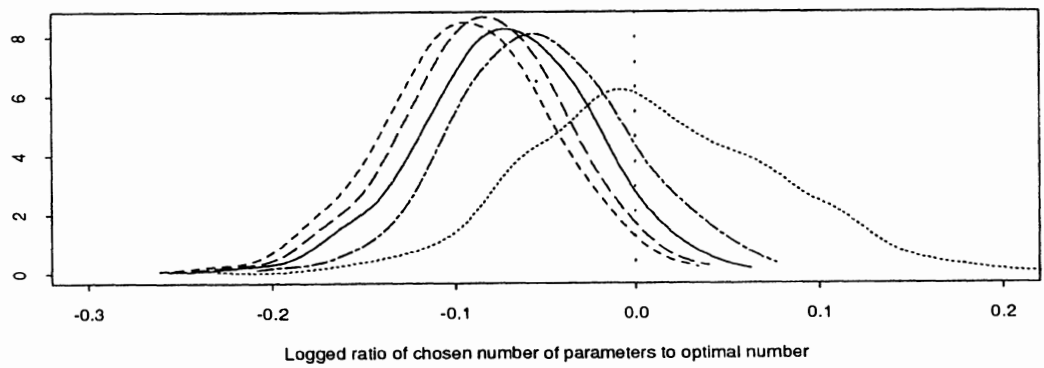


(c)

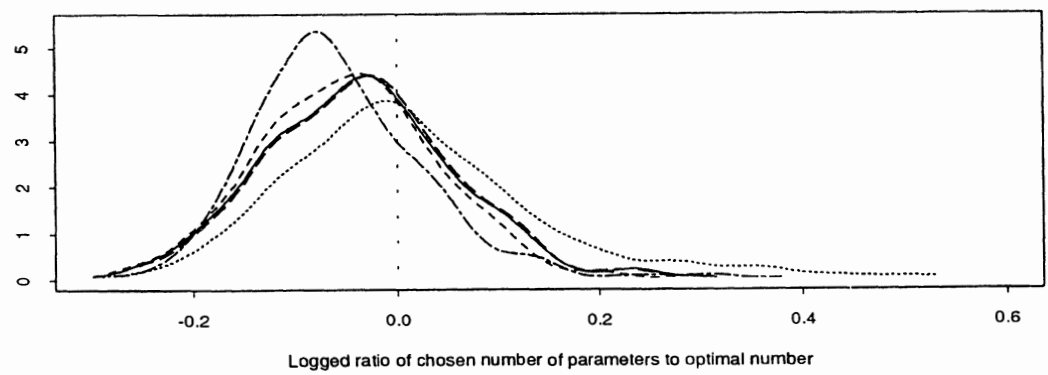
Fig. 3. Kernel density estimates of the distribution of logged ratios of the number of fitted parameters to the number corresponding to the minimizer of the ASE for the local quadratic estimator using AIC_C (———), GCV (·····), T (– – –), AIC_{C_1} (- - - -) and the plug-in method (— · —): (a) $m(x) = \sin(15\pi x)$, $\sigma/R_y = 0.01$; (b) $m(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$, $\sigma/R_y = 0.01$; (c) $m(x) = 10\ \exp(-10x)$, $\sigma/R_y = 0.05$



(a)

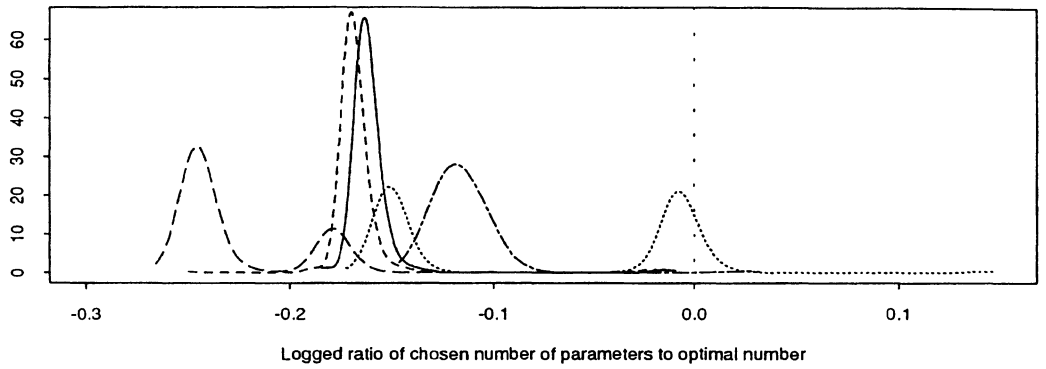


(b)

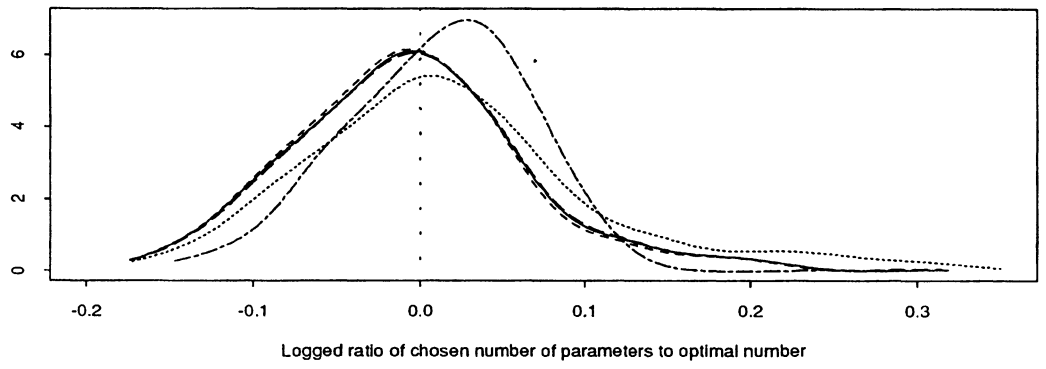


(c)

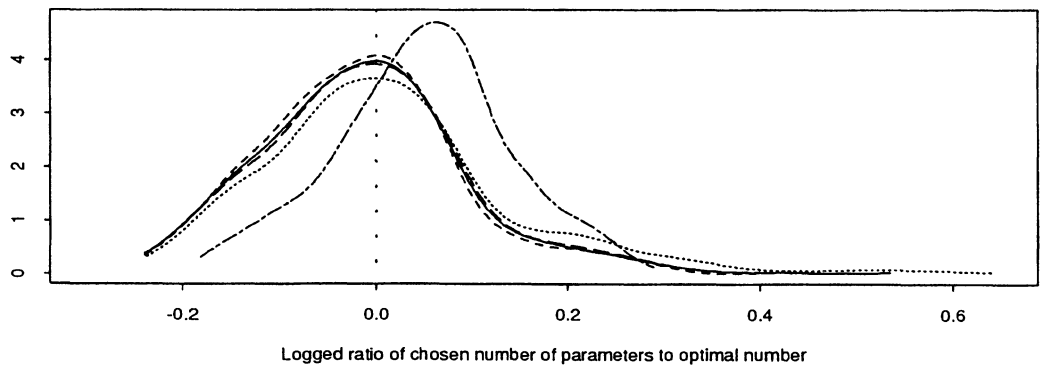
Fig. 4. Kernel density estimates of the distribution of logged ratios of the number of fitted parameters to the number corresponding to the minimizer of the ASE for the second-order convolution kernel estimator using AIC_C (———), GCV (·····), T (– – –), AIC_{C_1} (- - - -) and the plug-in method (— · —): (a) $m(x) = \sin(15\pi x)$, $\sigma/R_y = 0.01$; (b) $m(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$, $\sigma/R_y = 0.01$; (c) $m(x) = 10\exp(-10x)$, $\sigma/R_y = 0.05$



(a)

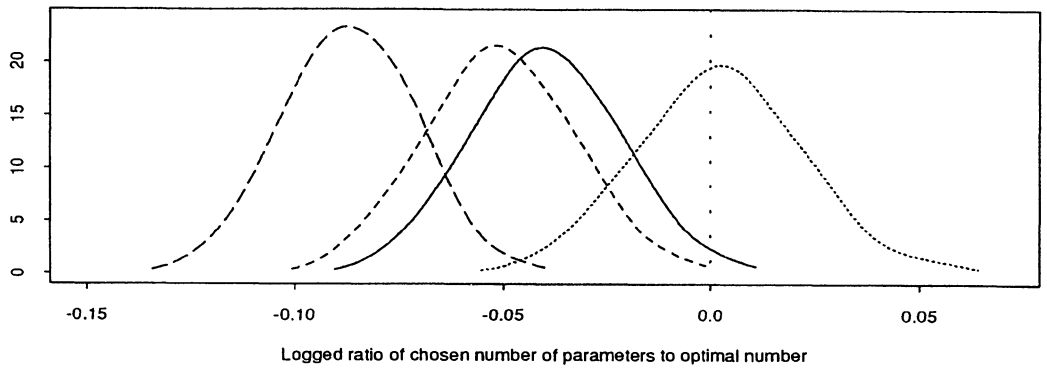


(b)

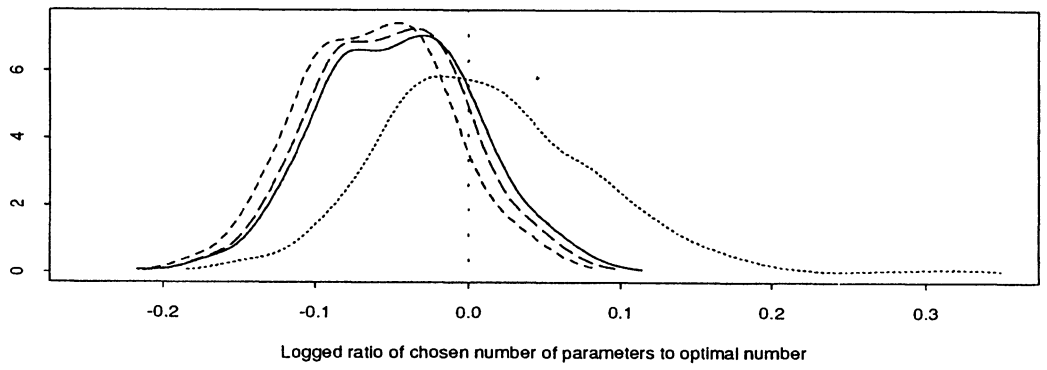


(c)

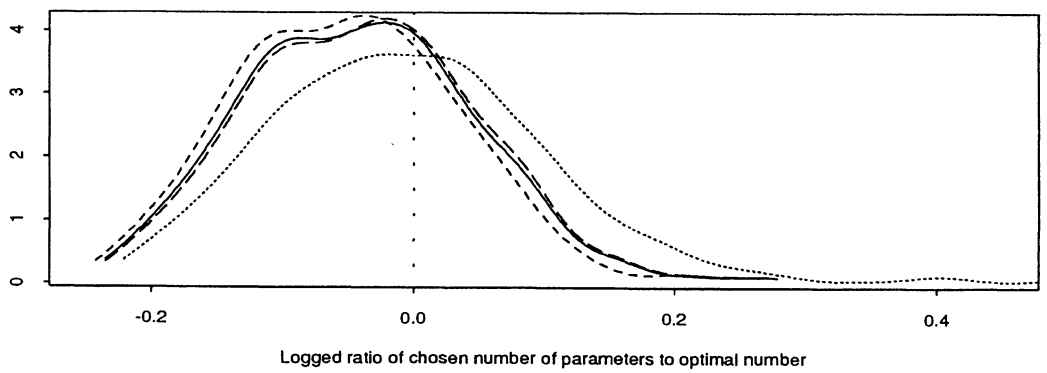
Fig. 5. Kernel density estimates of the distribution of logged ratios of the number of fitted parameters to the number corresponding to the minimizer of the ASE for the fourth-order convolution kernel estimator using AIC_C (———), GCV (·····), T (– – –), AIC_{C_1} (- - - -) and the plug-in method (— · —): (a) $m(x) = \sin(15\pi x)$, $\sigma/R_y = 0.01$; (b) $m(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$, $\sigma/R_y = 0.01$; (c) $m(x) = 10\exp(-10x)$, $\sigma/R_y = 0.05$



(a)



(b)



(c)

Fig. 6. Kernel density estimates of the distribution of logged ratios of the number of fitted parameters to the number corresponding to the minimizer of the ASE for the smoothing spline estimator using AIC_C (———), GCV (·····), T (– – –) and AIC_{C_1} (- - - -): (a) $m(x) = \sin(15\pi x)$, $\sigma/R_y = 0.01$; (b) $m(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$, $\sigma/R_y = 0.01$; (c) $m(x) = 10\exp(-10x)$, $\sigma/R_y = 0.05$

bandwidth for the local quadratic estimator that has the same asymptotic variance as the local linear estimator using $\hat{h}_L$ while having smaller asymptotic bias (Sheather, 1996). No results are given for the bandwidths using the AIC, since this rule almost invariably chose the smallest bandwidth tried in the simulation runs. The optimal and data-based bandwidths were found by using grid search routines. The results in Tables 1–4 will be discussed shortly.

Figs 2–6 show how the smoothing parameter selectors compare with the optimal choices for each of the regression estimators for three selected true regression functions (with an equispaced design and $n = 100$). The curves are kernel density estimates of the distribution of the logarithms (base 10) of the ratios of the effective number of fitted parameters chosen by the selector (i.e. $\mathrm{tr}(H)$) to the effective number of fitted parameters for the value that minimizes the ASE (i.e. the optimal value). By using this measure it is possible to compare the properties of the selectors for the different regression estimators, since the smoothing parameters themselves are not directly comparable. Note that average logged ratios greater than 0 (the right-hand side of the plots) correspond to undersmoothing, whereas average logged ratios that are less than 0 (the left-hand side of the plots) correspond to oversmoothing; this demarcation line is provided as an aid to interpretation.

Figs 2–6 describe properties for the three regression functions

(a) $m(x) = \sin(15\pi x)$, with $\sigma/R_y = 0.01$,
(b) $m(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$, with $\sigma/R_y = 0.01$, and
(c) $m(x) = 10\exp(-10x)$, with $\sigma/R_y = 0.05$.

These correspond to situations where either a relatively small, moderate or large amount of smoothing respectively is appropriate ((a) was deliberately chosen to represent an extreme low smoothing situation). Results are given for the local linear (Fig. 2), local quadratic (Fig. 3), second-order convolution kernel (Fig. 4), fourth-order kernel (Fig. 5) and cubic smoothing spline (Fig. 6) estimators.

The properties of the selectors differ for the various estimators, but certain patterns emerge. For all estimators, unless a very small amount of smoothing is appropriate, the curve for GCV has a noticeable long right-hand tail, corresponding to that selector's tendency to lead to undersmoothing. This is particularly apparent for the local polynomial estimators (Figs 2 and 3), where GCV cannot be considered to be sufficiently well behaved to use in practice. In virtually all situations the distribution of GCV is supported over a considerably wider range than the other selectors, reflecting its relatively high variability.

The $\mathrm{AIC_C}$-, $\mathrm{AIC_{C_1}}$- and T-selectors are generally similar to each other. They tend to oversmooth, although (except for local polynomial estimation and $m(x) = \sin(15\pi x)$) not dramatically so. They are considerably less variable than GCV in virtually all situations. Of these three selectors, $\mathrm{AIC_C}$ is usually best, in that it has the smallest tendency to oversmooth.

The properties of the plug-in selectors differ widely from estimator to estimator. All the plug-in selectors have relatively low variability, but their tendencies towards undersmoothing or oversmoothing are different. The local linear plug-in selector of Ruppert *et al.* (1995) does very poorly in the fine structure–strong regression situation (a), leading to severe oversmoothing (all the selectors except GCV strongly oversmooth, with $\mathrm{AIC_C}$ least extreme in this regard). Otherwise, its performance is quite reasonable.

Given its somewhat *ad hoc* nature, it is not surprising that the local linear-based local quadratic plug-in selector (Fig. 3) does not do well when compared with selectors designed to target the actual optimal bandwidth. It leads to oversmoothing in situation (a) and undersmoothing in situations (b) and (c). Its low variability does not make up for these difficulties, and $\mathrm{AIC_C}$, $\mathrm{AIC_{C_1}}$ and T are clearly better behaved.

The plug-in methods for the convolution kernel estimators (Figs 4 and 5), in contrast, seem to be comparatively well 'tuned' for these estimators in these situations. As noted earlier, they have low variability, although there is a small tendency towards oversmoothing for the second-order kernel and undersmoothing for the fourth-order kernel. Overall the plug-in selectors are probably the best choice for these estimators, with AIC_C being second best.

An interesting result is that the classical selectors perform better for the cubic smoothing spline (Fig. 6) than they do for the other estimators. GCV, for example, is noticeably less variable and somewhat less likely to undersmooth for this estimator, lending some support to its widespread use for smoothing splines in practice (it is still generally more variable and more likely to undersmooth than AIC_C is, however).

Tables 1–3 describe Monte Carlo results for an equispaced design and $n = 100$. These squared error results generally support the impressions given in Figs 2–6. Box plots of the ASEs for each simulation run (not given here) show that the summary measures given in Tables 1–3 do reflect the actual relative behaviour of the selectors, i.e. the mean ratio of ASE to optimal ASE does not reflect unusual values, but rather the actual pattern of the entire distribution of ASE values. The median ratios, though generally smaller than the mean ratios, follow the same patterns as the mean values. Signed rank tests comparing the paired ASE values for any two selectors are generally statistically significant at a 0.05 level if the difference in mean ASE ratios is greater than 0.02–0.05 for all the selectors except GCV; the high variability of this selector implies that differences in mean ASE ratio up to 0.15 are sometimes not statistically significant when the GCV selector is involved in the comparison.

Table 1 refers to the local linear estimator. The plug-in selector is most often best, and (except for $m(x) = \sin(15\pi x)$, where it fails badly) is usually not far from best otherwise. GCV does well when a small bandwidth is appropriate (although even then often worse than the plug-in estimator), but it deteriorates when a moderate or large bandwidth is best, because of its tendency to undersmooth. T, AIC_C and AIC_{C_1} are usually similarly behaved, with AIC_C noticeably better. Overall, AIC_C is competitive with the plug-in selector, though usually resulting in a slightly higher ASE.

The optimal mean ASE is uniformly lower for the local quadratic estimator (Table 2) compared with the local linear estimator, and for strong regression relationships (small σ/R_y) it is often 40–50% smaller. This clear superiority of the local quadratic estimator is consistent with the asymptotic properties and can be contrasted with the situation in kernel density estimation, where higher order kernel estimators only improve on second-order kernels for sample sizes in the hundreds and even thousands (Marron and Wand, 1992). The advantage of the local quadratic estimator lessens for weaker regression relationships and is small for regression function (vi), where the kink in the function means that the local quadratic estimator is not asymptotically superior to the local linear estimator, but in no cases is the local quadratic estimator worse than the local linear estimator. This is not true for the local cubic estimator, where Monte Carlo simulations (not given here) indicate that increased variability outweighs any advantages in boundary bias correction.

The second important point from Table 2 is that these available optimal gains from using the local quadratic estimator are achievable in practice. AIC_C is clearly the best choice, as it has good properties for all the situations examined. Once again GCV has problems when a large bandwidth is appropriate, whereas AIC_C is similar to, but consistently better than, T and AIC_{C_1}. The local linear-based local quadratic plug-in method beats the plug-in method applied to the local linear estimator, as it was designed to do, but it is not competitive with the other selectors under strong relationships for most of the regression functions. This is not surprising, since it is not targeting the actual optimal bandwidth for the local quadratic

Table 1. Monte Carlo results for the local linear estimator†

σ/R_y	*Results for the following estimators:*					
	Optimal	*GCV*	*T*	AIC_C	AIC_{C_1}	*Plug-in*
$m(x) = sin(15\pi x)$						
0.01	3.7238×10^{-4}	1.0718	17.4050	8.2626	15.0197	19.3300
0.05	5.5520×10^{-3}	1.8320	2.0377	1.4709	2.0328	2.0425
0.25	7.1133×10^{-2}	1.0808	1.1446	1.1119	1.2063	1.1571
0.5	0.21055	1.1261	1.1505	1.1574	1.2631	1.6895
$m(x) = sin(5\pi x)$						
0.01	1.7427×10^{-4}	1.0788	1.4794	1.2409	1.5125	1.0901
0.05	2.3635×10^{-3}	1.1101	1.0907	1.0782	1.1316	1.0300
0.25	3.0802×10^{-2}	1.1868	1.0926	1.0963	1.1032	1.0679
0.5	9.3359×10^{-2}	1.2736	1.1502	1.1541	1.1505	1.1138
$m(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$						
0.01	4.3702×10^{-4}	1.1124	1.1669	1.1172	1.1986	1.0503
0.05	5.7036×10^{-3}	1.1400	1.0770	1.0790	1.0984	1.0373
0.25	7.4161×10^{-2}	1.3431	1.1919	1.1933	1.1885	1.1837
0.5	0.22019	1.4465	1.2628	1.2570	1.2538	1.2720
$m(x) = 0.3\,exp\{-64(x-0.25)^2\} + 0.7\,exp\{-256(x-0.75)^2\}$						
0.01	2.4797×10^{-5}	1.9902	1.7656	1.3411	1.7844	1.2266
0.05	3.2843×10^{-4}	1.0867	1.1137	1.0901	1.1643	1.0911
0.25	4.1577×10^{-3}	1.1932	1.1085	1.1142	1.1287	1.0716
0.5	1.1892×10^{-2}	1.2550	1.1526	1.1605	1.1607	1.1301
$m(x) = 10\,exp(-10x)$						
0.01	2.3381×10^{-3}	1.1157	1.0993	1.0883	1.1429	1.0499
0.05	2.8251×10^{-2}	1.1676	1.0993	1.1042	1.1137	1.0614
0.25	0.34798	1.5630	1.3019	1.2982	1.2708	1.3356
0.5	1.0270	2.0272	1.5871	1.5708	1.5436	1.8056
$m(x) = exp(x - 1/3),\ x < 1/3;\ m(x) = exp\{-2(x - 1/3)\},\ x \geqslant 1/3$						
0.01	1.3842×10^{-5}	1.1096	1.1214	1.1067	1.1729	1.2017
0.05	1.6298×10^{-4}	1.1930	1.1055	1.1093	1.1130	1.0743
0.25	1.9694×10^{-3}	1.4431	1.2904	1.2847	1.2692	1.2819
0.5	5.7039×10^{-3}	2.1343	1.5719	1.5663	1.5220	1.6770

†$n = 100$. Entries are the average of the optimal ASEs and averages of ratios of the ASE to the optimal ASE.

estimator, but is only designed to beat the local linear estimator. Overall, Tables 1 and 2 clearly demonstrate that for $n = 100$ and an equispaced fixed design the best local polynomial choice is the local quadratic estimator using AIC_C to select the bandwidth.

Results for the second- and fourth-order convolution kernel estimators (not given here) are directly comparable with those in Tables 1 and 2 respectively. The optimal performance of the kernel estimators is comparable with that of the local linear and quadratic estimators, which is consistent with their asymptotic equivalence (i.e. the fourth-order kernel's optimal mean ASE is consistently smaller than that of the second-order kernel). The second-order plug-in selector is often better 'tuned' for the estimator than the plug-in estimator of Ruppert *et al.* (1995) is for the local linear estimator; the fourth-order plug-in selector is much better than the local linear-based local quadratic counterpart, since it targets the true optimal bandwidth. The effectiveness of GCV for small bandwidths combined with its ineffectiveness when large bandwidths are needed is again seen. Finally, AIC_C has the best overall

Table 2. Monte Carlo results for the local quadratic estimator†

σ/R_y	*Results for the following estimators:*					
	Optimal	*GCV*	*T*	AIC_C	AIC_{C_1}	*Plug-in*
$m(x) = sin(15\pi x)$						
0.01	2.0975×10^{-4}	1.1755	2.0879	1.1805	1.6248	6.5493
0.05	3.5698×10^{-3}	1.0617	1.2009	1.0821	1.1814	1.2433
0.25	6.1113×10^{-2}	1.0651	1.0827	1.0703	1.0992	1.1919
0.5	0.20298	1.0992	1.1608	1.1598	1.2449	1.8792
$m(x) = sin(5\pi x)$						
0.01	8.2870×10^{-5}	1.0883	1.0679	1.0633	1.0804	1.5036
0.05	1.4373×10^{-3}	1.1471	1.0715	1.0733	1.0754	1.3896
0.25	2.4541×10^{-2}	1.2279	1.1350	1.1368	1.1295	1.1819
0.5	8.3831×10^{-2}	1.2891	1.1712	1.1698	1.1631	1.1784
$m(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$						
0.01	2.0563×10^{-4}	1.1421	1.0828	1.0849	1.0875	1.6820
0.05	3.3589×10^{-3}	1.2552	1.1334	1.1326	1.1262	1.5152
0.25	5.8097×10^{-2}	1.3732	1.2920	1.2896	1.2797	1.4708
0.5	0.19511	1.4480	1.3180	1.3077	1.2923	1.4556
$m(x) = 0.3\,exp\{-64(x-0.25)^2\} + 0.7\,exp\{-256(x-0.75)^2\}$						
0.01	1.5243×10^{-5}	1.0658	1.1242	1.0683	1.1312	1.0545
0.05	2.6110×10^{-4}	1.0959	1.0675	1.0634	1.0819	1.0430
0.25	4.0800×10^{-3}	1.1812	1.1269	1.1328	1.1375	1.1053
0.5	1.2874×10^{-2}	1.2819	1.2018	1.2067	1.2070	1.1624
$m(x) = 10\,exp(-10x)$						
0.01	1.3335×10^{-3}	1.1596	1.1052	1.1080	1.1089	1.3682
0.05	1.9828×10^{-2}	1.2562	1.1790	1.1837	1.1727	1.3248
0.25	0.30346	1.7217	1.4073	1.4004	1.3888	1.5949
0.5	1.0119	1.6685	1.5791	1.5598	1.5497	1.9322
$m(x) = exp(x - 1/3),\ x < 1/3;\ m(x) = exp\{-2(x - 1/3)\},\ x \geq 1/3$						
0.01	1.2717×10^{-5}	1.1014	1.1066	1.1034	1.1288	1.1185
0.05	1.5444×10^{-4}	1.2179	1.1418	1.1446	1.1408	1.0855
0.25	1.9120×10^{-3}	1.4782	1.3590	1.3586	1.3526	1.3351
0.5	5.7491×10^{-3}	1.5550	1.4986	1.4991	1.4808	1.7313

†$n = 100$.

performance of the classical methods, being generally a little better than the plug-in estimator for larger target bandwidths and somewhat worse for smaller target bandwidths. The best single choice would probably be the plug-in selectors, but AIC_C is a reasonable alternative.

Table 3 summarizes results for the cubic smoothing spline estimator and $n = 100$. Asymptotically this estimator is comparable with the local quadratic and fourth-order kernel estimators, and the results bear this out, but only to a certain extent. Whereas for most regression functions the optimal mean ASE is close to that for the local quadratic and fourth-order kernel estimators, it is as much as 50% worse for functions (iii) and (v), which lack very fine structure. There is no plug-in selector for the smoothing spline, but all the classical selectors generally work at least as well (compared with the optimal performance) as they do for the other estimators. This is particularly true for GCV, but even so AIC_C has the best overall performance of all the selectors, except in the case of very fine structure (although the performances of T, AIC_C and AIC_{C_1} are generally very similar). Still, the comparatively poor performance of the spline estimator compared with the local quadratic and fourth-order

Table 3. Monte Carlo results for the cubic smoothing spline estimator†

σ/R_y	*Results for the following estimators:*				
	Optimal	*GCV*	*T*	AIC_C	AIC_{C_1}
$m(x) = sin(15\pi x)$					
0.01	1.9775×10^{-4}	1.0210	1.6827	1.1500	1.1935
0.05	3.4964×10^{-3}	1.0424	1.1665	1.0796	1.1409
0.25	6.0982×10^{-2}	1.0534	1.0832	1.0675	1.1052
0.5	0.20250	1.1074	1.2334	1.2575	1.3837
$m(x) = sin(5\pi x)$					
0.01	8.2714×10^{-5}	1.0656	1.0620	1.0557	1.0779
0.05	1.4395×10^{-3}	1.1075	1.0652	1.0661	1.0722
0.25	2.4680×10^{-2}	1.1428	1.1171	1.1183	1.1176
0.5	8.2995×10^{-2}	1.2540	1.1543	1.1548	1.1492
$m(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$					
0.01	3.1636×10^{-4}	1.0785	1.0851	1.0780	1.1080
0.05	4.5184×10^{-3}	1.1364	1.0968	1.1007	1.1014
0.25	6.6864×10^{-2}	1.2711	1.2042	1.2046	1.1939
0.5	0.20513	1.3900	1.2719	1.2657	1.2565
$m(x) = 0.3\,exp\{-64(x-0.25)^2\} + 0.7\,exp\{-256(x-0.75)^2\}$					
0.01	1.4642×10^{-5}	1.0447	1.1056	1.0638	1.1112
0.05	2.5468×10^{-4}	1.0652	1.0686	1.0631	1.0881
0.25	3.9257×10^{-3}	1.1652	1.1273	1.1330	1.1425
0.5	1.1705×10^{-2}	1.2553	1.2017	1.2167	1.2273
$m(x) = 10\,exp(-10x)$					
0.01	1.8821×10^{-3}	1.0975	1.0835	1.0830	1.1043
0.05	2.5465×10^{-2}	1.1675	1.1261	1.1315	1.1328
0.25	0.34545	1.4831	1.3128	1.3082	1.2973
0.5	1.0373	1.9060	1.6007	1.5909	1.5689
$m(x) = exp(x-1/3),\ x < 1/3;\ m(x) = exp\{-2(x-1/3)\},\ x \geqslant 1/3$					
0.01	1.2380×10^{-5}	1.0896	1.1085	1.1046	1.1400
0.05	1.4496×10^{-4}	1.2204	1.1419	1.1461	1.1425
0.25	1.7391×10^{-3}	1.4468	1.3573	1.3527	1.3415
0.5	5.3246×10^{-3}	1.9124	1.6375	1.6221	1.5845

†$n = 100$.

kernel estimators for the two regression functions mentioned above casts doubt on its use over the other estimators for an equispaced design.

Results for $n = 50$ and $n = 500$ (not given here) can be summarized as follows. For the smaller sample size, variability of the estimators is naturally a problem, and accordingly the low variability of the plug-in selectors works in their favour. For this reason the plug-in selectors are generally better than the other selectors, although AIC_C is still competitive. The plug-in selector's properties are very similar to those of the classical selectors (except GCV) when $n = 500$. The results of Hall and Johnstone (1992) imply that plug-in methods are asymptotically superior to most classical methods in terms of the ASE (having the same convergence rate, but a smaller constant).

It is well known that the convolution kernel (1.1) is asymptotically inefficient for random designs; for example, the asymptotic conditional variance of the optimal second-order kernel estimator is 1.5 times the conditional variance of the asymptotically optimal local linear estimator. Herrmann (1996) discussed general versions of the Gasser–Müller estimator

$$\hat{m}(x) = h^{-1} \sum_{i=1}^{n} c_i \left\{ \int_{a_i}^{b_i} K\left(\frac{x-u}{h}\right) \mathrm{d}u \right\} y_i, \tag{3.1}$$

pointing out that it is the variability of the differences $b_i - a_i$ ($(x_{i+1} - x_{i-1})/2$ for estimator (1.1)) that potentially inflates the variance of estimator (3.1) (see also Chu and Marron (1991) and Jones *et al.* (1994)). Herrmann suggested taking $c_i = 1$ and using kernel quantile estimators to determine a_i and b_i,

$$a_i = g^{-1} \sum_{j=1}^{n} x_j \int_{(j-0.5)/(n+1)}^{(j+0.5)/(n+1)} K_s \left\{ \frac{(i-0.5)/(n+1) - v}{g} \right\} \mathrm{d}v$$

and

$$b_i = g^{-1} \sum_{j=1}^{n} x_j \int_{(j-0.5)/(n+1)}^{(j+0.5)/(n+1)} K_s \left\{ \frac{(i+0.5)/(n+1) - v}{g} \right\} \mathrm{d}v,$$

where K_s is a symmetric boundary-corrected kernel of order $k_s = 2$ or $k_s = 4$,

$$g = 0.75(n+1)^{-(3k+1)/(2k+1)(k_s+1)},$$

and k is the order of the kernel K. This estimator does not suffer the asymptotic inefficiency under random designs of estimator (1.1) (a different approach to this problem is given by Hall and Turlach (1997)).

Simulation results when the predictor values in each simulation run were taken as a random sample from a uniform distribution (not given here) were similar to those for a fixed uniform grid. The mean ASE values for the local quadratic and cubic smoothing spline estimators were close to those for the fixed uniform design, which is consistent with those estimators' asymptotic equivalence under fixed and random designs. The values for Herrmann's modified convolution kernel estimator were sometimes considerably higher, however, indicating that the corrective action has not taken hold at $n = 100$. The plug-in selector for this estimator also did not approach as close to the optimal ASE as the fixed uniform design version does in some situations. The $\mathrm{AIC_C}$ selector, in contrast, generally performed well.

Table 4 gives representative results ($n = 100$) for when the predictor values fall in a fixed non-uniform grid. The values satisfy $x_i = \exp(i/20 - 5)$, $i = 1, \ldots, 100$, yielding a set of values that are much more tightly packed at the low end than at the high end (x_{39} is closer to 0 than x_{99} is to 1). In this situation a fixed local polynomial or kernel bandwidth is not optimal, but the performance of fixed bandwidth selectors is still of interest. No values are given for the local linear-based local quadratic plug-in bandwidth because the algorithm frequently did not converge to an answer. The results are consistent with those for fixed and random uniform designs, in that

(a) GCV works well for small σ, but less well for large σ,
(b) the convolution kernel plug-in selector sometimes does not approach as close to the optimal ASE as the fixed uniform design version does and
(c) the $\mathrm{AIC_C}$ selector generally performs well in all circumstances.

A simple way to allow the bandwidth of the local quadratic estimator to vary with the design density is to base it on a fixed number of nearest neighbours rather than on a fixed

Table 4. Monte Carlo results for a non-uniform fixed design†

σ/R_y	*Results for the following estimators:*					
	Optimal	*GCV*	*T*	AIC_C	AIC_{C_1}	*Plug-in*
Local quadratic estimator, $m(x) = \sin(5\pi x)$						
0.01	8.7407×10^{-5}	1.0670	1.0666	1.0635	1.0733	
0.05	1.4579×10^{-3}	1.1087	1.0813	1.0830	1.0839	
0.25	2.4684×10^{-2}	1.2098	1.1505	1.1519	1.1523	
0.5	8.3624×10^{-2}	1.2943	1.2235	1.2240	1.2135	
Fourth-order kernel estimator, $m(x) = \sin(5\pi x)$						
0.01	1.0683×10^{-4}	1.1824	1.2084	1.2037	1.2179	1.6998
0.05	1.5433×10^{-3}	1.1514	1.1116	1.1131	1.1132	1.3636
0.25	2.3408×10^{-2}	1.3557	1.2896	1.2908	1.2820	1.2661
0.5	7.3234×10^{-2}	1.4176	1.3008	1.3048	1.3026	1.2826
Cubic smoothing spline estimator, $m(x) = \sin(5\pi x)$						
0.01	8.2293×10^{-5}	1.0647	1.0605	1.0562	1.0735	
0.05	1.4260×10^{-3}	1.1233	1.0683	1.0697	1.0747	
0.25	2.4355×10^{-2}	1.1470	1.1153	1.1165	1.1159	
0.5	8.1981×10^{-2}	1.2382	1.1445	1.1452	1.1413	
Local quadratic estimator, $m(x) = 0.3\exp\{-64(x-0.25)^2\} + 0.7\exp\{-256(x-0.75)^2\}$						
0.01	1.1687×10^{-5}	1.0575	1.0721	1.0619	1.0706	
0.05	2.1593×10^{-4}	1.0924	1.0727	1.0751	1.0838	
0.25	3.2169×10^{-3}	1.2119	1.1795	1.1852	1.1862	
0.5	9.5174×10^{-3}	1.4309	1.3247	1.3290	1.3191	
Fourth-order kernel estimator, $m(x) = 0.3\exp\{-64(x-0.25)^2\} + 0.7\exp\{-256(x-0.75)^2\}$						
0.01	1.4170×10^{-5}	1.0479	1.0567	1.0534	1.0534	2.0179
0.05	2.2239×10^{-4}	1.1446	1.0769	1.0815	1.0832	1.2631
0.25	3.2546×10^{-3}	1.2659	1.2035	1.2051	1.2074	1.1628
0.5	9.7789×10^{-3}	1.4871	1.3454	1.3373	1.3365	1.2305
Cubic smoothing spline estimator, $m(x) = 0.3\exp\{-64(x-0.25)^2\} + 0.7\exp\{-256(x-0.75)^2\}$						
0.01	1.4242×10^{-5}	1.0435	1.0977	1.0652	1.0996	
0.05	2.5024×10^{-4}	1.0628	1.0660	1.0617	1.0830	
0.25	3.8233×10^{-3}	1.1611	1.1260	1.1311	1.1387	
0.5	1.1463×10^{-2}	1.2603	1.2243	1.2345	1.2462	

†$n = 100$.

distance, as is done in most implementations of LOESS (this distinction does not matter for uniform designs except in the boundary region). Choosing the number of nearest neighbours by using GCV, T, AIC_C and AIC_{C_1} was investigated here as well. It turned out, however, that the optimal ASE achieved by the local quadratic LOESS estimator was consistently at least 25% larger than that achieved by the fixed bandwidth local quadratic estimator, and sometimes more than three times larger, so this estimator cannot be recommended on the basis of an ASE criterion.

We conclude this section with a real data example. Fig. 7 refers to the data set, which relates the concentration of nitric oxide in engine exhaust (normalized by engine work) to the equivalence ratio, a measure of the richness of the air–ethanol mix, for burning ethanol in a single-cylinder automobile test engine (Brinkman, 1981). Local quadratic estimates based on AIC_C (full curve, $\hat{h} = 0.0382$), GCV (dotted curve, $\hat{h} = 0.0227$) and a plug-in method based on the local linear plug-in estimator (broken curve, $\hat{h} = 0.0426$) are superimposed on the plot. GCV leads to undersmoothing, whereas AIC_C gives a very reasonable representation of the

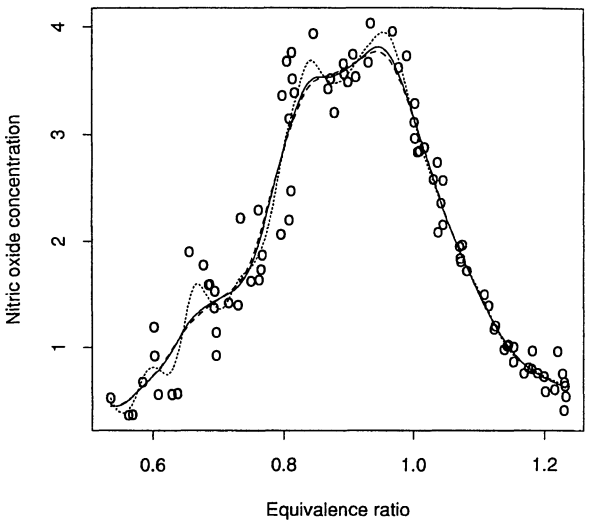


Fig. 7. Ethanol data: superimposed are local quadratic estimates with smoothing parameters chosen by using AIC_C (———), GCV (·····) and the local linear-based plug-in method (- - - -)

data (including the bimodality in the centre of the curve). The local linear-based local quadratic plug-in estimate is seemingly slightly oversmoothed but is generally similar to the AIC_C-based estimate.

The fitted GCV- and AIC_C-based regression estimates in Fig. 7 for these data do not give a complete impression of the properties of these selectors for these data. The GCV criterion is very flat for these data over a wide range of bandwidths (differing by less than 0.5% of its value over the range $\hat{h} \in [0.0185, 0.04]$), while being minimized towards the low end of this interval. AIC_C, in contrast, is much less flat over the range of bandwidths considered. These results are consistent with the pattern seen in the Monte Carlo simulations, since the flatness of the GCV criterion is consistent with high variability, and the minimum at small h is consistent with the tendency to undersmooth.

4. Conclusions

The construction of effective smoothing parameters for nonparametric regression estimators has been a source of much research, and much controversy, in recent years. In this paper we have proposed a variant of the AIC that avoids some of the difficulties of other selectors, and that can be applied easily for use with any linear smoother.

The results here leave open several problems. The local polynomial and kernel estimators can be modified so that the bandwidth can be locally varied to give better estimates by accounting for local differences in curvature of m, density f_X and variance $\sigma^2(x)$. Several researchers have proposed automatic selection methods for this, and it would be interesting to see whether AIC_C could also be applied to this important problem.

An application to smoothing estimators based on principles other than least squares would also be valuable. Just as generalized linear models (McCullagh and Nelder, 1989) generalize regression models to binomial, Poisson and other data types, so also can smoothing methods be generalized to other data types through the likelihood function. Recent examples of such methods are given in Gu and Qiu (1994) and Fan *et al.* (1995). The Poisson regression model

provides a natural link to local likelihood density estimation (Hjort and Jones, 1996; Loader, 1996) through categorical data smoothing (Simonoff (1996), chapter 6), implying potential applications to these other smoothing problems.

Our finding that AIC_C tends to undersmooth less than GCV parallels an analogous result for model selection in parametric linear regression. In Hurvich (1997) it is shown that in the parametric case AIC_C is guaranteed to select a model which is at least as parsimonious as that selected by using Tukey's MS/ν criterion. The MS/ν criterion (discussed in Anscombe (1967), Tukey (1967) and Mosteller and Tukey (1977), p. 386) is defined as the ratio of the residual sum of squares to the square of the residual degrees of freedom and therefore is the exact parametric analogue of the GCV criterion. Thus, for parametric models, AIC_C tends to select a model with fewer parameters than does the analogue of GCV, yielding a less undersmoothed estimate of the mean function.

Acknowledgements

The authors would like to thank Eva Herrmann, Matt Wand and the Joint Editor for helpful discussions of this material, and Eva and Matt for providing relevant computer code. Chih-Ling Tsai's research was supported in part by National Science Foundation grant DMS-95-10511.

Appendix A: S-PLUS functions

S-PLUS functions and the data set used in Section 3 can be obtained as an S-PLUS dump file using the World Wide Web at the location

```
http://www.blackwellpublishers.co.uk/rss
```

These functions can be used to determine the AIC_C-based smoothing parameters for the nearest neighbour local polynomial estimator LOESS (`aicc.loess`) and the cubic smoothing spline (`aicc.spline`) respectively. The functions determine the minimizer of AIC_C by using the function minimizer `nlminb` (rather than using a grid search, as was done in the results summarized in Section 3). Each function takes the predictor vector `x` and response vector `y` as input and returns as output the appropriate smoother object (`loess` or `smooth.spline` respectively) and the value of AIC_C.

It is possible that the function minimizer will not find the true minimum, and for this reason we suggest that the functions are run several times with the starting values `start` set to different values (this is particularly important if there is a relatively small number of distinct predictor values, since then the criterion will have many local plateaus). The criteria functions (`crit.loess` and `crit.spline`) can also be used to perform a grid search over the appropriate range of smoothing parameters to determine the true minimizer of AIC_C.

A Postscript file of an expanded version of this paper is available via the World Wide Web at the location

```
http://www.stern.nyu.edu/~jsimonof/aicc.ps
```

References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Information Theory* (eds B. N. Petrov and F. Csàki), pp. 267–281. Budapest: Akadémiai Kiadó.

———(1974) A new look at the statistical model identification. *IEEE Trans. Autom. Control*, **19**, 716–723.

Anscombe, F. J. (1967) Topics in the investigation of linear relations fitted by the method of least squares (with discussion). *J. R. Statist. Soc.* B, **29**, 1–52.

Brinkman, N. D. (1981) Ethanol fuel—a single cylinder engine study of efficiency and exhaust emissions. *SAE Trans.*, **90**, 1410–1427.

Chu, C.-K. and Marron, J. S. (1991) Choosing a kernel regression estimator (with discussion). *Statist. Sci.*, **6**, 404–436.

Cleveland, W. S. and Devlin, S. J. (1988) Locally weighted regression: an approach to regression analysis by local fitting. *J. Am. Statist. Ass.*, **83**, 596–610.
Craven, P. and Wahba, G. (1979) Smoothing noisy data with spline functions. *Numer. Math.*, **31**, 377–403.
Fan, J., Heckman, N. E. and Wand, M. P. (1995) Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Am. Statist. Ass.*, **90**, 141–150.
Gasser, T. and Müller, H. G. (1979) Kernel estimation of regression functions. *Lect. Notes Math.*, **757**, 23–68.
Grund, B., Hall, P. and Marron, J. S. (1994) Loss and risk in smoothing parameter selection. *J. Nonparam. Statist.*, **4**, 107–132.
Gu, C. and Qiu, C. (1994) Penalized likelihood regression: a simple asymptotic analysis. *Statist. Sin.*, **4**, 297–304.
Hall, P. and Johnstone, I. (1992) Empirical functionals and efficient smoothing parameter selection (with discussion). *J. R. Statist. Soc.* B, **54**, 475–530.
Hall, P. and Marron, J. S. (1991) Lower bounds for bandwidth selection in density estimation. *Probab. Theory Reltd Flds*, **90**, 149–173.
Hall, P. and Turlach, B. A. (1997) Enhancing convolution and interpolation methods for nonparametric regression. *Biometrika*, **84**, in the press.
Härdle, W., Hall, P. and Marron, J. S. (1988) How far are automatically chosen regression smoothing parameters from their optimum? *J. Am. Statist. Ass.*, **83**, 86–101.
Hart, J. D. and Yi, S. (1996) One-sided cross-validation. Unpublished.
Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. London: Chapman and Hall.
Herrmann, E. (1996) On the convolution type kernel regression estimator. Unpublished.
———(1997) Local bandwidth choice in kernel regression estimation. *J. Comput. Graph. Statist.*, **6**, 35–54.
Hjort, N. L. and Jones, M. C. (1996) Locally parametric density estimation. *Ann. Statist.*, **24**, 1619–1647.
Hurvich, C. M. (1997) Mean square over degrees of freedom: new perspectives on a model selection treasure. In *The Practice of Data Analysis, in Honor of John W. Tukey* (eds D. Brillinger, L. T. Fernholz and S. Morgenthaler). Princeton: Princeton University Press. To be published.
Hurvich, C. M. and Tsai, C.-L. (1989) Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.
Jones, M. C. (1986) Expressions for inverse moments of positive quadratic forms in normal variables. *Aust. J. Statist.*, **28**, 242–250.
———(1987) On moments of ratios of quadratic forms in normal variables. *Statist. Probab. Lett.*, **6**, 129–136.
———(1991) The roles of ISE and MISE in density estimation. *Statist. Probab. Lett.*, **12**, 51–56.
Jones, M. C., Davies, S. J. and Park, B. U. (1994) Versions of kernel-type regression estimators. *J. Am. Statist. Ass.*, **89**, 825–832.
Jones, M. C. and Kappenman, R. F. (1991) On a class of kernel density estimate bandwidth selectors. *Scand. J. Statist.*, **19**, 337–349.
Khatri, C. G. (1980) Quadratic forms in normal variables. In *Handbook of Statistics*, vol. 1 (ed. P. R. Krishnaiah), pp. 443–469. Amsterdam: North-Holland.
Kotz, S. and Johnson, N. L. (eds) (1986) *Encyclopedia of Statistical Sciences*, vol. 7. New York: Wiley.
Linhart, H. and Zucchini, W. (1986) *Model Selection*. New York: Wiley.
Loader, C. R. (1995) Old Faithful erupts: bandwidth selection reviewed. Unpublished.
———(1996) Local likelihood density estimation. *Ann. Statist.*, **24**, 1602–1618.
Mammen, E. (1990) A short note on optimal bandwidth selection for kernel estimators. *Statist. Probab. Lett.*, **9**, 23–25.
Marron, J. S. and Wand, M. P. (1992) Exact mean integrated squared error. *Ann. Statist.*, **20**, 712–736.
McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
Mosteller, F. and Tukey, J. W. (1977) *Data Analysis and Regression*. Reading: Addison-Wesley.
Rice, J. (1984) Bandwidth choice for nonparametric regression. *Ann. Statist.*, **12**, 1215–1230.
Ruppert, D., Sheather, S. J. and Wand, M. P. (1995) An effective bandwidth selector for local least squares regression. *J. Am. Statist. Ass.*, **90**, 1257–1270.
Sheather, S. J. (1996) Bandwidth selection: plug in methods versus classical methods. *Joint Statistical Meet., Chicago*.
Simonoff, J. S. (1996) *Smoothing Methods in Statistics*. New York: Springer.
Terrell, G. R. (1992) Discussion of 'The performance of six popular bandwidth selection methods on some real data sets' and 'Practical performance of several data driven bandwidth selectors'. *Comput. Statist.*, **7**, 275–277.
Tukey, J. W. (1967) Discussion on Topics in the investigation of linear relations fitted by the method of least squares (by F. J. Anscombe). *J. R. Statist. Soc.* B, **29**, 47–48.
Turlach, B. A. and Wand, M. P. (1996) Fast computation of auxiliary quantities in local polynomial regression. *J. Comput. Graph. Statist.*, **5**, 337–350.