Computational Statistics (1999) 14:515–532

Bandwidth selection for local linear regression: A simulation study

Thomas C. M. Lee[1] and Victor Solo[2]

[1] The University of Chicago, Department of Statistics, 5734 S. University Avenue, Chicago, Illinois 60637, U.S.A.
[2] Department of Statistics, Macquarie University, Sydney, NSW 2109, Australia

Summary

This paper provides a simulation study of several popular bandwidth selectors for local linear regression. The study also includes two new selectors which couple the non–asymptotic plug–in and the unbiased risk estimation techniques. These two new selectors are simple to describe, easy to implement and performed very well in our simulation study.

Keywords: Bandwidth selection, Least–square cross–validation, Local linear regression, Plug–in, PURE

[1] Corresponding author.

1 Introduction

In the past decade, local polynomial regression has emerged as an attractive procedure for nonparametric function estimation. Particularly in the case of random X's, it has superior bias and variance properties (e.g., Fan & Gijbels 1996, Section 3.2.1) as well as avoiding boundary bias problems suffered by other methods such as various versions of kernel estimators (e.g., Fan & Gijbels 1996, Section 3.2.5).

The bandwidth of course has a crucial effect on the quality of the estimator, and a number of methods have been developed for the automatic choice of bandwidth. A natural bandwidth selector is least–squares cross–validation (LSCV); for example, see Fan & Gijbels (1996, Section 4.10.2). In addition, various methods based on the plug–in idea have also been proposed. Gasser, Kneip & Köhler (1991) and Ruppert, Sheather & Wand (1995) developed plug–in bandwidth selection methods which rely on some *asymptotic* expressions (the former was developed for a different nonparametric estimator, but applies also to local linear regression), while Fan & Gijbels (1995) developed some plug–in bandwidth selection methods which rely on *non–asymptotic* expressions.

The aim of this paper is twofold. First, we introduce a new method which can be applied to construct bandwidth selectors. This method is based on a coupling of the non–asymptotic Plug–in and the Unbiased Risk Estimation techniques. We call our method PURE and our work was inspired by the work of Chiu (1992), as we describe further below.

The second goal of the paper is to evaluate the practical performance of our new method as well as those above–mentioned existing bandwidth selection methods by means of a simulation study.

The rest of this paper is organized as follows. Section 2 provides some background of local linear regression. Section 3 surveys four existing bandwidth selectors while Section 4 presents our PURE methodology. Results of a simulation study are reported in Section 5. Section 6 discusses a special application of our PURE methodology, and Section 7 concludes this paper. In the sequel we write $\boldsymbol{X}_1^n = \{X_1, \ldots, X_n\}$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$.

2 Local Linear Regression

Suppose we observe n pairs of measurements (X_i, Y_i), $i = 1, \ldots, n$, relating to the model

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

where ϵ_i is an iid $(0, \sigma^2)$ noise while $m(X)$ is an unknown function of interest. We suppose X_i is drawn at random from an unknown sampling density $p(x)$ with a known compact support $[a, b]$. The local linear regression estimator $\hat{m}_h(x)$ is obtained by fitting local straight lines (e.g., see Wand & Jones 1995, Chapter 5 or Fan & Gijbels 1996, Chapter 1) and is given by

$$\hat{m}_h(x) = w_{x,h}^T Y,$$

where

$$\begin{aligned} w_{x,h} &= (w_{x,h}(X_1), \ldots, w_{x,h}(X_n))^T, \\ w_{x,h}(X_i) &= \frac{\{s_{2,x,h} - s_{1,x,h}(X_i - x)\} K_h(X_i - x)}{n \{s_{2,x,h} s_{0,x,h} - (s_{1,x,h})^2\}}, \\ s_{r,x,h} &= \sum_{i=1}^{n} (X_i - x)^r K_h(X_i - x). \end{aligned}$$

Here h is the bandwidth and $K(\cdot)$ is a nonnegative kernel obeying the standard conditions: $\int K(u)du = 1$, $\int uK(u)du = 0$ and $0 < \mu_2 = \int u^2 K(u)du < \infty$. The function $K_h(u)$ is defined as $K_h(u) = \frac{1}{h}K(\frac{u}{h})$. The bandwidth h controls the size of "features" or "bumps" in the unknown function that maybe discerned from the available noisy data. It might better be called a characteristic length or resolution width rather than a bandwidth.

Amongst others, there are two legitimate measures of estimator quality, namely, the mean integrated squared error (MISE) and the weighted mean integrated squared error (WMISE). Note that the latter sometimes is called the predictive integrated squared error and that the *mean* in both MISE and WMISE is taken over the distribution of ϵ_i. Also, there are both conditional and unconditional versions of these two measures. The conditional versions are

$$\begin{aligned} \mathrm{MISE}_c &= E\left[\int \{m(x) - \hat{m}_h(x)\}^2 dx | \boldsymbol{X}_1^n\right], \\ \mathrm{WMISE}_c &= E\left[\int \{m(x) - \hat{m}_h(x)\}^2 p(x) dx | \boldsymbol{X}_1^n\right] \end{aligned}$$

while the unconditional versions are

$$\mathrm{MISE}_u = E(\mathrm{MISE}_c) \qquad \text{and} \qquad \mathrm{WMISE}_u = E(\mathrm{WMISE}_c).$$

Notice that all these four measures are unknown quantities as m is unknown. Most existing methods aim at minimizing one of the two conditional measures (MISE_c or WMISE_c). The primary concern of this paper is also the conditional versions, but we also discuss how to apply our PURE idea to construct bandwidth selectors which aim to minimize either MISE_u or WMISE_u.

A remark: there are conflicting views regarding the question of whether one should consider the unconditional versions at all (e.g., see Jones 1991 for such a discussion in the density estimation context). In order to allow the readers to form their own views, we shall keep both the conditional and unconditional versions in our paper.

To aid with the description of some existing bandwidth selectors we note the asymptotic expression for WMISE_c (e.g., Ruppert et al., 1995)

$$\text{AWMISE}_c = \frac{h^4\mu_2^2}{4}\int \{m''(x)\}^2 p(x)dx + \frac{\sigma^2 R(K)(b-a)}{nh} \tag{1}$$

where $R(K) = \int K^2(u)du$. From this expression we find the asymptotically optimal bandwidth for minimizing AWMISE_c is

$$\begin{aligned} h_{\text{AWMISE}} &= \left\{ \frac{\sigma^2 R(K)(b-a)}{n\mu_2^2 \int \{m''(x)\}^2 p(x)dx} \right\}^{\frac{1}{5}} \\ &= C_1(K)\left\{\frac{\sigma^2(b-a)}{n\theta_{22}}\right\}^{\frac{1}{5}}, \end{aligned} \tag{2}$$

where $C_1(K) = \{R(K)/\mu_2^2\}^{\frac{1}{5}}$ and for later use we have introduced the array

$$\theta_{rs} = \int m^{(r)}(x)m^{(s)}(x)p(x)dx, \qquad r,s \geq 0, \qquad r+s \text{ even.}$$

Again for later use we write down the following estimator, with a bandwidth g, for σ^2:

$$\hat{\sigma}_g^2 = \frac{1}{\nu}\sum_{i=1}^{n}\{Y_i - \hat{m}_g(X_i)\}^2. \tag{3}$$

This is simply a normalized residual sum of squares and the normalizing quantity ν, sometimes known as the degrees of freedom, is given by $\nu = n - 2\sum_i w_{X_i,h}(X_i) + \sum_{i,j} w_{X_j,h}^2(X_i)$. Its presence guarantees $E(\hat{\sigma}_g^2|X_1,\ldots,X_n) = \sigma^2$ whenever $m(x)$ is either constant or linear.

3 Existing Bandwidth Selection Methods

In this section we briefly survey four existing bandwidth selection methods.

Least–Squares Cross–Validation:. Perhaps the most well–known bandwidth selection method is least–squares cross–validation. It selects the bandwidth h that minimizes the cross–validation score

$$\text{CV}(h) = \sum_{i=1}^{n}\{Y_i - \hat{m}_{h,-i}(X_i)\}^2, \tag{4}$$

where $\hat{m}_{h,-i}(X_i)$ is the usual "leave-one-out" estimate of $m(X_i)$.

Asymptotic Direct Plug–In: Ruppert et al. (1995) developed a "direct plug–in" bandwidth selection procedure which relies on the asymptotically optimal bandwidth expression (2). The idea is to plug–in estimates of σ^2 and θ_{22} into expression (2), and obtain an estimate of h_{AWMISE} as the chosen bandwidth.

Asymptotic Iterative Plug–In: An "iterative plug–in" bandwidth selection procedure for the Gasser–Müller kernel estimator (Gasser & Müller 1984) was developed by Gasser et al. (1991). This iterative plug–in scheme can also be applied to local linear regression. As for the direct plug–in approach, it relies on the asymptotic expression (2); i.e., it also aims to minimize WMISE_c. However, the asymptotic optimal bandwidth h_{AWMISE} is estimated in an iterative manner.

Non–Asymptotic Plug–In: Fan & Gijbels (1995) developed some bandwidth selection procedures which also involve the plug–in notion. However, there is a major difference between their plug–in procedures and the two asymptotic procedures mentioned above: in their procedures the plug–in idea is only applied to *non–asymptotic* expressions. The particular procedure developed by Fan & Gijbels that we are interested here is the so–called global refined bandwidth selector. However, in this paper we shall call it non–asymptotic plug–in.

4 PURE Bandwidth Selectors

In this section we present two new PURE bandwidth selectors, one aims to minimize MISE_c while the other aims to minimize WMISE_c. We begin with the MISE_c. It is straightforward to show the bias–variance decomposition

$$\text{MISE}_c = \int \left\{m(x) - \boldsymbol{w}_{x,h}^T \boldsymbol{m}\right\}^2 dx + \sigma^2 \int \boldsymbol{w}_{x,h}^T \boldsymbol{w}_{x,h} dx.$$

One way to construct an estimator of MISE_c is to plug–in pilot local linear estimates $\hat{m}_g(x)$ and $\hat{\sigma}_g^2$ (see (3)) into the above expression, and obtain

$$\widehat{\text{MISE}}_c = \int \left\{\hat{m}_g(x) - \boldsymbol{w}_{x,h}^T \hat{\boldsymbol{m}}_g\right\}^2 dx + \hat{\sigma}_g^2 \int \boldsymbol{w}_{x,h}^T \boldsymbol{w}_{x,h} dx,$$

where $\hat{\boldsymbol{m}}_g = (\hat{m}_g(X_1), \ldots, \hat{m}_g(X_n))^T$. For simplicity we suggest choosing the pilot bandwidth g by LSCV, and we use simple quadrature to calculate the integrals. Our PURE bandwidth selector which aims to minimize MISE_c chooses the bandwidth as the minimizer of $\widehat{\text{MISE}}_c$.

Similarly we can estimate WMISE_c by

$$\int \{\hat{m}_g(x) - \boldsymbol{w}_{x,h}^T \hat{\boldsymbol{m}}_g\}^2 p(x)dx + \hat{\sigma}_g^2 \int \boldsymbol{w}_{x,h}^T \boldsymbol{w}_{x,h} p(x)dx.$$

Now the first term is still a problem since we do not know $p(x)$. A second use of the plug–in principle gives the estimator

$$\widehat{\text{WMISE}}_c = \frac{1}{n}\sum_{j=1}^{n}\left\{\hat{m}_g(X_j) - \boldsymbol{w}_{X_j,h}^T \hat{\boldsymbol{m}}_g\right\}^2 + \hat{\sigma}_g^2 \int \boldsymbol{w}_{x,h}^T \boldsymbol{w}_{x,h} p(x)dx.$$

Our PURE bandwidth selector which aims to minimize WMISE_c chooses the bandwidth as the minimizer of $\widehat{\text{WMISE}}_c$.

Thus, the general idea of our PURE methodology is that, we first obtain an *exact* and *suitable* expression (see below) for MISE_c (or WMISE_c), and then apply the *non–asymptotic* plug–in principle to obtain an estimate of MISE_c (or WMISE_c). For simplicity we recommend using LSCV to obtain the relevant pilot bandwidth(s). Our numerical experience suggests that LSCV is very reasonable for pilot bandwidth selection.

In the above when we said an "*exact* and *suitable* expression" for MISE_c, we meant the expression should be non–asymptotic and should not contain "higher order" quantities like m''. As a result, our two PURE bandwidth selectors, unlike other plug–in bandwidth selectors, do not require the existence of m''.

We have the following remarks.

1. The PURE methodology can be easily extended to higher–order local polynomials.

2. Our procedure was inspired by the "double CV" method of Chiu (1992) for kernel density estimation. Chiu (1992) also estimates terms in an exact MISE formula. He obtains a simple pilot density estimator in which the pilot bandwidth (actually a cutoff frequency) is chosen by CV. This pilot estimator is plugged into the exact MISE formula, and the bandwidth for the main density estimator is chosen by minimizing the resulting MISE expression. The procedure is also described by Hall, Marron & Park (1992) and called smoothed CV by them. We rather see it as a smoothed unbiased risk estimation procedure.

3. There are several features of Chiu's method that we think are important. Firstly it is easy to describe and simple to implement — just two density estimators are used and two bandwidths are chosen (no asymptotic expressions are needed). Further the method has the potential to be applied in more complex bandwidth selection problems for inverse problems, such as nonlinear ones, where asymptotic expressions may

not be available. Our procedure then has the same simplicity and ease of implementation as Chiu's method since no asymptotic expressions are needed.

4. Chiu (1991) also discusses nonparametric function estimation but only for regularly gridded data. Also he does not give an automatic method to choose the pilot bandwidth. Thus we see our work as a natural extension of his density estimation procedure.

5. A referee pointed out that our PURE methodology is closely related to the "double smoothing" methods described in Härdle, Hall & Marron (1992) and Wand & Gutierrez (1997). However we have pushed the simple idea of using LSCV to choose the pilot, and backed up our suggestion by numerical results to be reported in the next section. We would also like to mention that our work was done independently.

6. To facilitate the following discussion, we first distinguish *pilot* function estimate from *final* function estimate: for the current PURE context the *pilot* function estimate is the LSCV local linear regression estimate $\hat{m}_g$ (see expression for $\widehat{\text{MISE}}_c$ or $\widehat{\text{WMISE}}_c$), while the *final* function estimate is the local linear regression estimate obtained using the minimizer of $\widehat{\text{MISE}}_c$ or $\widehat{\text{WMISE}}_c$ as bandwidth. Work of Härdle et al. (1992) on kernel regression suggests that an optimal asymptotic MISE rate of $1/\sqrt{n}$ cannot be achieved if both the pilot and final function estimates are of the same structure. This issue can be dealt with by using higher order kernels in the pilot estimation stage (a point missed by Härdle et al. 1992). The equivalent change in local polynomial regression is to use higher order polynomials for pilot estimation (e.g, see Fan & Gijbels 1996). We have repeated our simulation study using a pilot function estimate with local cubic, and obtained very similar results, and hence we do not report them here.

5 Simulation Study

5.1 Settings

A simulation study was conducted to evaluate the practical performance of the proposed bandwidth selectors, as well as the four existing bandwidth selectors discussed in Section 3. Four Test Functions were used:

$$\begin{array}{lcr} 1: & \sin(10\pi x), & x \in [0,1], \\ 2: & 1-48x+218x^2-315x^3+145x^4, & x \in [0,1], \\ 3: & \sin(2x)+2\exp(-16x^2), & x \in [-2,2], \\ 4: & x+2\exp(-16x^2), & x \in [-2,2]. \end{array}$$

Test Functions 1 and 2 were used by Ruppert et al. (1995) while Test Functions 3 and 4 were used by Fan & Gijbels (1995). Also, Test Functions 1 and 4 are of the same structural form as two of the test functions used by Gasser et al. (1991). These four functions, together with their Fourier representations, are plotted in Fig. 1.

Three signal–to–noise ratios (snrs) and two design densities were used. That is, each Test Function was tested with six combinations of snr and design density. Here snr is defined to be the variance of the function divided by the variance of the noise: snr $= \mathrm{var}(m)/\sigma^2$. The three snrs were: low = 2, medium = 4 and high = 8, and the two design densities were the uniform density and $p_0(x) = 6x(1-x)I_{\{0 \leq x \leq 1\}}$ (for Test Functions 3 and 4 we used an adjusted version of $p_0(x)$).

The design density $p_0(x)$ was used by Hastie & Loader (1993). Note that $p_0(x)$ provides a sparse design near the boundaries, and that results for $p_0(x)$ cannot be generalized to other non–uniform random designs such as one that is having positive limits at boundaries.

The kernel that we used was Gaussian, hence we did not have a lot of problems with data sparseness. But if sparseness did occur for a particular simulated data set, we replaced it with a new set (we only needed to replace less than 5% of data sets). One could also use the interpolation methods of Hall & Turlach (1997) to handle this data sparseness problem.

For each combination of Test Function, snr and design density, 200 sets of simulated data were generated, and the number of data points n for each data set was 50. We used one n because n and snr are interchangeable, as the (asymptotic) performance depends only on the ratio σ^2/n. This simple observation does not seem to have been emphasized before.

For each of these generated data sets, we computed the following automatically selected bandwidths:

h_{PUREM}	:	PURE bandwidth selector which aims to minimize MISE_c,
h_{PUREW}	:	PURE bandwidth selector which aims to minimize WMISE_c,
h_{LSCV}	:	least–squares cross–validation bandwidth selector,
h_{DAPI}	:	direct asymptotic plug–in bandwidth selector,
h_{IAPI}	:	iterative asymptotic plug–in bandwidth selector,
h_{NAPI}	:	non–asymptotic plug–in bandwidth selector.

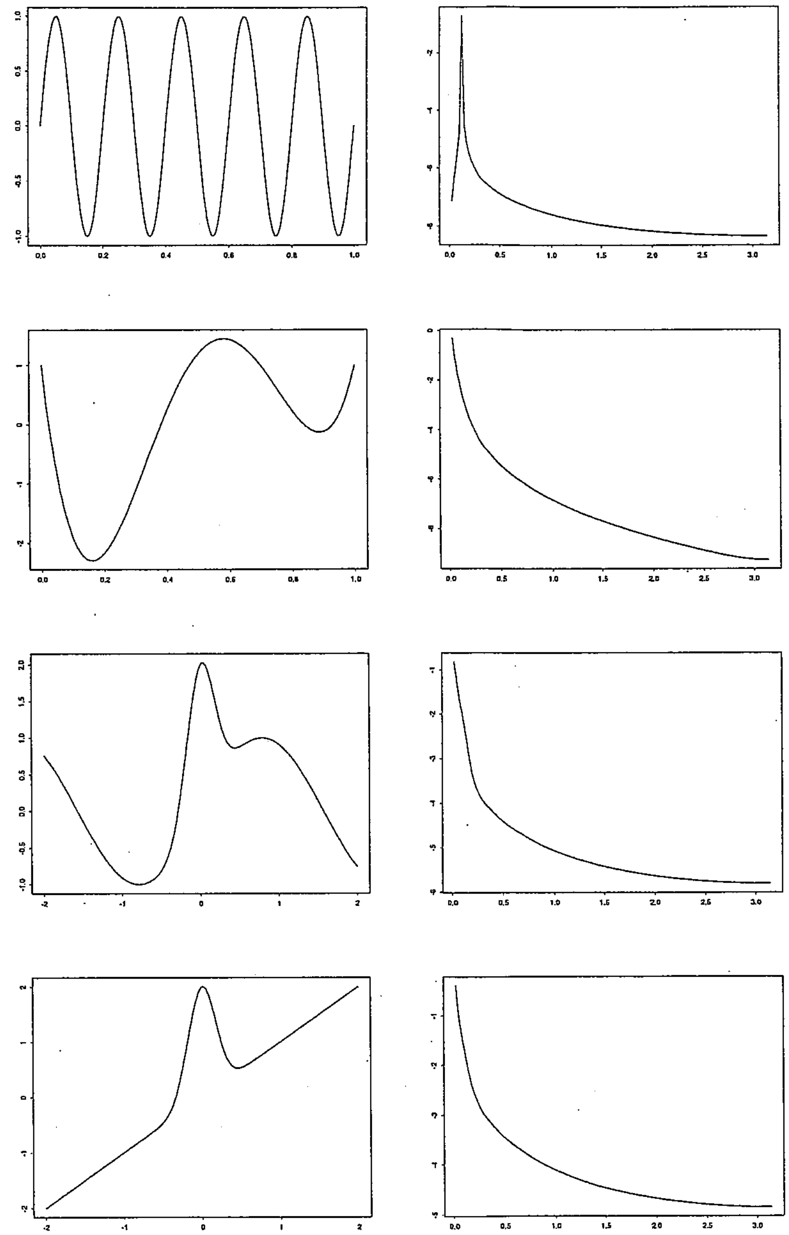

Figure 1: Plots of Test Functions. Plots in the left column, from top to bottom, are Test Functions 1 to 4 respectively. Plots in the right column are the corresponding Fourier representations, plotted in a log scale. Notice that for Test Function 1, its Fourier representation is a delta function.

5.2 Results: Uniform Random Design

For the uniform design, in addition to the above six automatically computed bandwidths, we also computed the optimal but unknown bandwidth h_{ISE}, defined as the minimizer of a discretized version of

$$\text{ISE}(h) = \int \{m(x) - \hat{m}_h(x)\}^2 \, dx$$

(note for uniform design ISE = WISE; see Section 5.3 for definition of WISE).

We only present results corresponding to median snr, as results for high and low snrs are similar.

Fig. 2 displays the boxplots of the bandwidth ratios

$$(h - h_{\text{ISE}})/h_{\text{ISE}},$$

where h is any of the six automatically selected bandwidths. Also, boxplots of the ISE ratios

$$\log\{\text{ISE}(h)/\text{ISE}(h_{\text{ISE}})\}$$

are given in Fig. 3.

For Test Function 1, both h_{DAPI} and h_{IAPI} had a strong tendency to oversmooth. In fact, for some simulated data sets, both h_{DAPI} and h_{IAPI} were far too large to the extent that the obvious sinusoidal structure of Test Function 1 was completely smoothed out!

For Test Function 2, h_{NAPI} performed slightly worse than the others. It seems that h_{NAPI} tended to undersmooth.

For Test Functions 3 and 4, all bandwidth selectors gave similar performance, with the slight exception that, h_{IAPI} sometimes produced a large ISE ratio.

Perhaps the most interesting observation is that, for most cases the performance of h_{LSCV} was actually better than h_{DAPI} and h_{IAPI}. We will return to this observation later.

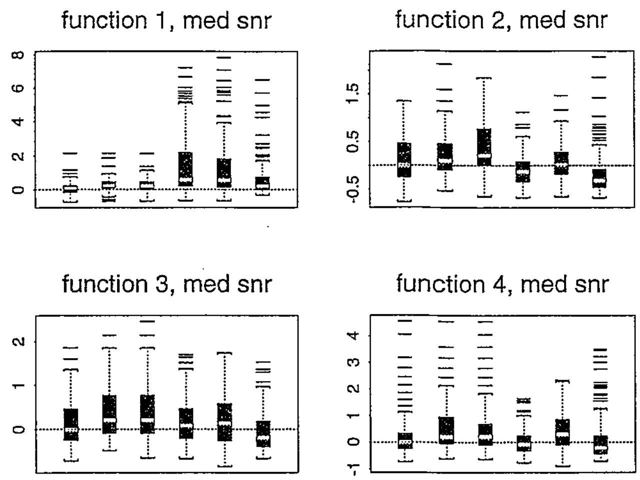

Figure 2: Boxplots of $(h - h_{\text{ISE}})/h_{\text{ISE}}$ for various h, uniform random design and medium snr. In each panel, the boxplots correspond to (from left to right) h_{PUREM}, h_{PUREW}, h_{LSCV}, h_{DAPI}, h_{IAPI} and h_{NAPI}.

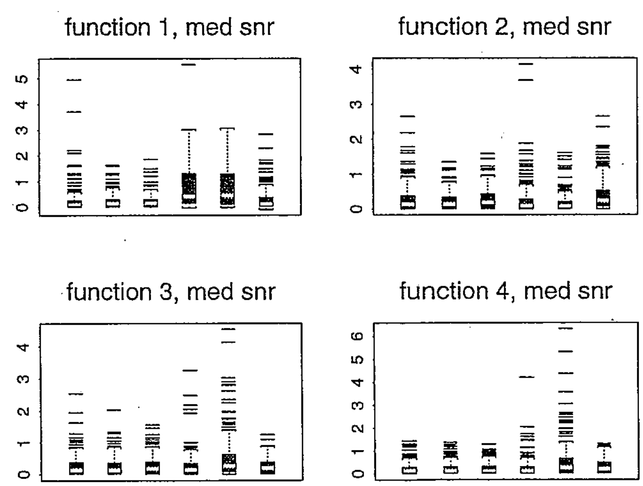

Figure 3: Boxplots of $\log\{\text{ISE}(h)/\text{ISE}(h_{\text{ISE}})\}$ for various h, uniform random design and medium snr. In each panel, the boxplots correspond to (from left to right) h_{PUREM}, h_{PUREW}, h_{LSCV}, h_{DAPI}, h_{IAPI} and h_{NAPI}.

5.3 Results: Non–Uniform Random Design

Here we report our results when the design density was $p_0(x)$. Again, we only report results corresponding to medium snr.

For each of the 200 simulated data sets, we computed the six automatically selected bandwidths, the optimal but unknown bandwidth h_{ISE}, and the optimal but unknown bandwidth h_{WISE} which minimizes a discretized version of

$$\text{WISE}(h) = \int \{m(x) - \hat{m}_h(x)\}^2 \, p(x) dx.$$

That is, eight bandwidths were computed.

Figs. 4 and 5 display the boxplots of

$$(h - h_{\text{ISE}})/h_{\text{ISE}} \qquad \text{and} \qquad \log\{\text{ISE}(h)/\text{ISE}(h_{\text{ISE}})\}$$

respectively, while Figs. 6 and 7 display the boxplots of

$$(h - h_{\text{WISE}})/h_{\text{WISE}} \qquad \text{and} \qquad \log\{\text{WISE}(h)/\text{WISE}(h_{\text{WISE}})\}$$

respectively. Here h is any of the six automatically selected bandwidths.

In general, the relative ranking of the six bandwidth selectors is roughly the same in terms of both ISE(h) and WISE(h). However, it seems that h_{IAPI} was inferior to other selectors, while h_{DAPI} performed well if WISE(h) is the criterion for measuring quality. For h_{NAPI}, it performed well for Test Function 4, but poorly for Test Function 1. For other cases, we feel that the performance of the bandwidth selectors was similar.

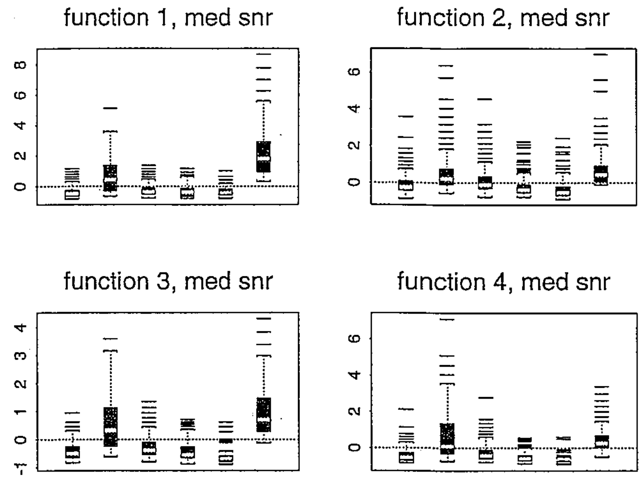

Figure 4: Boxplots of $(h - h_{\text{ISE}})/h_{\text{ISE}}$ for various h with design density $p_0(x)$ and medium snr. In each panel, the boxplots correspond to (from left to right) h_{PUREM}, h_{PUREW}, h_{LSCV}, h_{DAPI}, h_{IAPI} and h_{NAPI}.

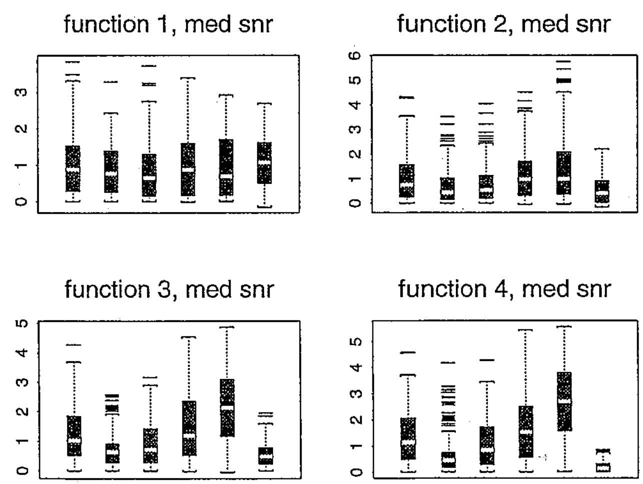

Figure 5: Boxplots of $\log\{\text{ISE}(h)/\text{ISE}(h_{\text{ISE}})\}$ for various h with design density $p_0(x)$ and medium snr. In each panel, the boxplots correspond to (from left to right) h_{PUREM}, h_{PUREW}, h_{LSCV}, h_{DAPI}, h_{IAPI} and h_{NAPI}.

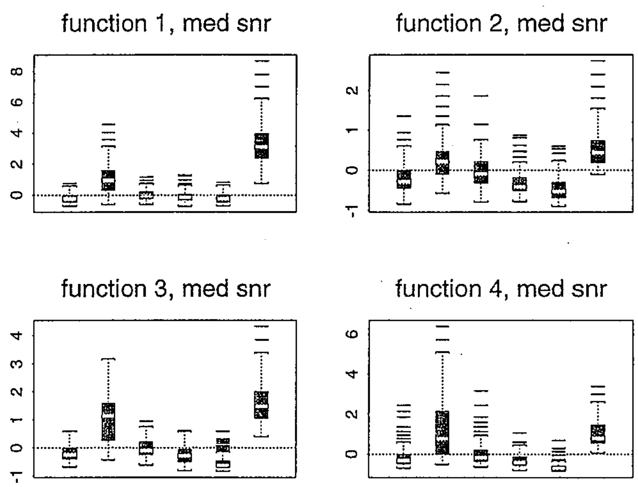

Figure 6: Boxplots of $(h - h_{\mathrm{WISE}})/h_{\mathrm{WISE}}$ for various h with design density $p_0(x)$ and medium snr. In each panel, the boxplots correspond to (from left to right) h_{PUREM}, h_{PUREW}, h_{LSCV}, h_{DAPI}, h_{IAPI} and h_{NAPI}.

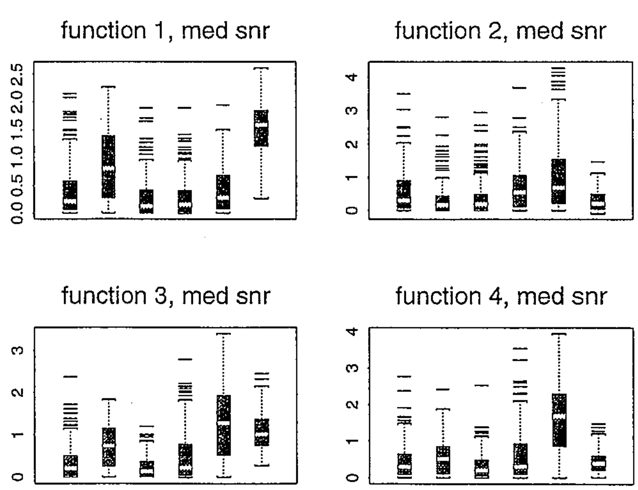

Figure 7: Boxplots of $\log\{\mathrm{WISE}(h)/\mathrm{WISE}(h_{\mathrm{WISE}})\}$ for various h with design density $p_0(x)$ and medium snr. In each panel, the boxplots correspond to (from left to right) h_{PUREM}, h_{PUREW}, h_{LSCV}, h_{DAPI}, h_{IAPI} and h_{NAPI}.

5.4 Empirical Conclusion

No bandwidth selector performed uniformly the best in our simulation. However, there are a few important empirical conclusions:

1. the selectors h_{DAPI} and h_{IAPI} sometimes ridiculously oversmooth,
2. the selectors h_{DAPI}, h_{IAPI} and h_{NAPI}, which have been shown to have faster rate of convergence than h_{LSCV}, do not outperform h_{LSCV} in our study, and
3. the two new selectors h_{PUREM} and h_{PUREW}, even though not always the best, compare favorably with other selectors in virtually all cases.

6 Unconditional Versions

In this section we briefly discuss how to adopt our PURE methodology to construct a bandwidth selector which aims to minimize the unconditional version of MISE, i.e., MISE_u. The case of WMISE can be tackled in a similar manner, and hence is omitted.

We first express MISE_u in a form so that the PURE methodology can be directly applied. Recall that $\text{MISE}_u = E(\text{MISE}_c)$ and let $\boldsymbol{X}_1^{*n} = \{X_1^*, \ldots, X_n^*\}$ be n independent draws from $p(x)$. Direct calculations give

$$\text{MISE}_u = E^* \left[\int \left\{ m(x) - \boldsymbol{w}_{x,h}^{*T} \boldsymbol{m}^* \right\}^2 dx \right] + \sigma^2 \int E^* \left(\boldsymbol{w}_{x,h}^{*T} \boldsymbol{w}_{x,h}^* \right) dx,$$

where $\boldsymbol{m}^* = (m(X_1^*), \ldots, m(X_n^*))^T$ and $\boldsymbol{w}_{x,h}^* = (w_{x,h}(X_1^*), \ldots, w_{x,h}(X_n^*))^T$. Now we need to obtain an estimator of MISE_u, and this can be achieved by using the bootstrap.

Let $\boldsymbol{X}_1^{bn} = \{X_1^b, \ldots, X_n^b\}$ be the bth bootstrap sample obtained from the original observed X_i's. That is, each of the X_i^b's is drawn from $\{X_1, \ldots, X_n\}$ with replacement. A natural estimator of MISE_u is:

$$\widehat{\text{MISE}}_u = \frac{1}{B} \sum_{b=1}^{B} \int \left\{ m(x) - \boldsymbol{w}_{x,h}^{bT} \boldsymbol{m}^b \right\}^2 dx + \hat{\sigma}_g^2 \frac{1}{B} \sum_{b=1}^{B} \int \boldsymbol{w}_{x,h}^{bT} \boldsymbol{w}_{x,h}^b dx,$$

where $\boldsymbol{m}^b$ and $\boldsymbol{w}_{x,h}^b$ are defined similarly to $\boldsymbol{m}^*$ and $\boldsymbol{w}_{x,h}^*$ (with $\boldsymbol{X}_1^{*n}$ replaced by $\boldsymbol{X}_1^{bn}$), and B is the number of bootstrap replicates. Thus, if the aim is to minimize MISE_u, we can choose the bandwidth as the minimizer of $\widehat{\text{MISE}}_u$.

However, this bootstrap or unconditional PURE approach has one major drawback: it is extremely computational intensive. It is B times slower than the non–bootstrap or conditional PURE.

Efron & Tibshirani (1993, Chapter 18) discuss another bootstrap procedure for estimating WMISE suited to this context. Their procedure is more elegant in that the pilot bandwidth estimation is not needed. However, their procedure also has the same drawback as the bootstrap PURE procedure: it is extremely computational intensive.

To study the performance of the bootstrap PURE procedure and the bootstrap procedure of Efron & Tibshirani, we performed a small experiment. We generated 50 data sets from Test Function 1 with snr=4 and a uniform design density. Again, the number of data points for each data set was 50, but we used a different error measure, PSE(h) instead of ISE(h):

$$\mathrm{PSE}(h) = \sum_{i=1}^{200} \{y_i^* - \hat{m}_h(x_i^*)\}^2,$$

where (x_i^*, y_i^*), $i = 1, \ldots, 200$, are new simulated "future" data.

For each generated data set, we computed five bandwidths: 1. h_{PSE} (the optimal bandwidth which minimizes PSE); 2. h_{PUREM} (non–bootstrap PURE); 3. h_{LSCV} (least–square cross–validation); 4. h_{BOOT} (bootstrap PURE); and 5. h_{ET} (Efron & Tibshirani's selector).

The results are displayed in a similar fashion as before: Fig. 8 displays the boxplots of the bandwidth ratios $(h - h_{\mathrm{PSE}})/h_{\mathrm{PSE}}$, and the boxplots of the PSE ratios $\log\{\mathrm{PSE}(h)/\mathrm{PSE}(h_{\mathrm{PSE}})\}$. Here h is any of the four automatically selected bandwidths. The results seem to suggest that the extra effort of bootstrapping does not improve the performance. However, this statement is far from conclusive.

7 Discussion

In this paper we surveyed four existing bandwidth selectors and proposed two new bandwidth selectors for local linear regression. All six selectors were empirically assessed by means of a simulation study. Numerical results demonstrate that the two proposed selectors compare favorably with other selectors.

Numerical results also suggest that, some plug–in bandwidth selectors appearing in the literature are in fact inferior to the LSCV selector, even though these plug–in selectors have been shown to possess better *theoretical* properties than the LSCV selector. This observation agrees with the study reported in Chiu (1996) in the context of density estimation.

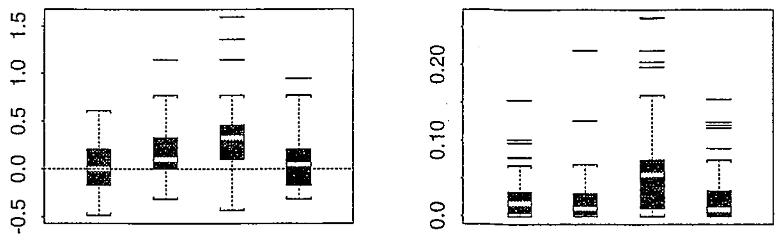

Figure 8: Results concerning the bootstrap–based procedures. Left: boxplots of $(h - h_{\text{PSE}})/h_{\text{PSE}}$ for various h. Right: boxplots of $\log\{\text{PSE}(h)/\text{PSE}(h_{\text{PSE}})\}$ for various h. In each panel, the boxplots correspond to (from left to right) h_{LSCV}, h_{PUREM}, h_{BOOT} and h_{ET}.

The two new bandwidth selectors are based on the method of coupling the non–asymptotic plug–in and the unbiased risk estimation techniques. This method is very easy to use and requires little theory to explain. Also, this method can be applied to more complex problems whereas asymptotic plug–in rely on analytic expressions for bias, variance and optimal bandwidth which will not always be available. Theoretical properties of our method will be described elsewhere, but one can expect results similar to those of Chiu (1992).

Acknowledgment

Most of the work was done while the first author was with the Department of Statistics, Macquarie University. The authors thank Steve Marron, three anonymous referees and an associate editor for many helpful comments, and Matt Wand for providing a preprint of Wand & Gutierrez (1997). The work was partially supported by an Australian Research Council grant. This document was prepared using computer facilities supported in part by National Science Foundation Grant DMS 89-05292 awarded to the Department of Statistics at The University of Chicago, and by The University of Chicago Block Fund.

References

Chiu, S.-T. (1991), 'Some stabilized bandwidth selectors for nonparametric regression', *The Annals of Statistics* **19**, 1528–1546.

Chiu, S.-T. (1992), 'An automatic bandwidth selector for kernel density estimation', *Biometrika* **79**, 771–782.

Chiu, S.-T. (1996), 'A comparative review of bandwidth selection for kernel density estimation', *Statistica Sinica* **6**, 129–145.

Efron, B. & Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Fan, J. & Gijbels, I. (1995), 'Data–driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation', *Journal of the Royal Statistical Society Series B* **57**, 371–394.

Fan, J. & Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, Chapman and Hall, London.

Gasser, T. & Müller, H. G. (1984), 'Estimating regression functions and their derivatives by the kernel method', *Scandinavian Journal of Statistics* **11**, 171–185.

Gasser, T., Kneip, A. & Köhler, W. (1991), 'A flexible and fast method for automatic smoothing', *Journal of the American Statistical Association* **86**, 643–652.

Hall, P. & Turlach, B. A. (1997), 'Interpolation methods for adapting to sparse design in nonparametric regression', *Journal of the American Statistical Association* **92**, 466–476.

Hall, P., Marron, J. S. & Park, B. U. (1992), 'Smoothed cross–validation', *Probability Theory and Related Fields* **92**, 1–20.

Hastie, T. & Loader, C. (1993), 'Local regression: automatic kernel carpentry (with discussion)', *Statistical Sciences* **8**, 120–143.

Härdle, W., Hall, P. & Marron, J. S. (1992), 'Regression smoothing parameters that are not far from their optimum', *Journal of the American Statistical Association* **87**, 227–233.

Jones, M. C. (1991), 'The roles of ISE and MISE in density estimation', *Statistics and Probability Letters* **12**, 51–56.

Ruppert, D., Sheather, S. J. & Wand, M. P. (1995), 'An effective bandwidth selector for local least squares regression', *Journal of the American Statistical Association* **90**, 1257–1270.

Wand, M. & Gutierrez, R. G. (1997), 'Exact risk approaches to smoothing parameter selection', *Journal of Nonparametric Statistics* **8**, 337–354.

Wand, M. P. & Jones, M. C. (1995), *Kernel Smoothing*, Chapman and Hall, London.