

Automatic Smoothing Parameter Selection:

A Survey

by

J. S. Marron

University of North Carolina

January 25, 1988

## ABSTRACT

This is a survey of recent developments in smoothing parameter selection for curve estimation. The first goal of this paper is to provide an introduction to the methods available, with discussion at both a practical and also a nontechnical theoretical level, including comparison of methods. The second goal is to provide access to the literature, especially on smoothing parameter selection, but also on curve estimation in general. The two main settings considered here are nonparametric regression and probability density estimation, although the points made apply to other settings as well. These points also apply to many different estimators, although the focus is on kernel estimators, because they are the most easily understood and motivated, and have been at the heart of the development in the field.

## 1. Introduction

Choice of smoothing parameter is the central issue in the application of all types of nonparametric curve estimators. This is demonstrated in Figure 1 which shows a simulated regression setup. In Figure 1a, the curve is the underlying regression function, and simulated observations, taken at equally spaced design points, are represented by crosses. Figures 1b, 1c and 1d show the same curve and observations together with some moving weighted averages of the crosses, shown as dashed curves, corresponding to different window widths, as shown by the dashed curves representing the weights, which appear at the bottom of each plot. Note that in Figure 1b, the window width is quite narrow, with the result that there are not enough observations appearing in each window for stability of the average, and the resulting estimate is overly subject to sample variability, i.e. is too "wiggly". Note that this is improved in Figure 1c, where a larger window width has been used. In Figure 1d, the window width is so large that observations from too far away appear in the averages, with the effect of introducing some bias, or in other words features of the underlying curve that are actually present have been smoothed away.

[put figure 1 about here]

The very large amount of flexibility in nonparametric curve estimators, demonstrated by changing the window width in the example of Figure 1, allows great payoffs, because these estimators do not arbitrarily impose structure on the data, which is always done by parametric estimators. To see how this is the case, think of doing a

simple linear regression least squares fit of the data in Figure 1. Of course, if the structure imparted by a parametric model is appropriate, then that model should certainly be used for inference, as the decreased flexibility allows for much firmer results, in terms of more powerful hypothesis test, and smaller confidence intervals. However it is in cases where no model readily suggests itself, or there may be some doubt as to the model, that nonparametric curve estimation really comes to the fore. See Silverman (1986) and Härdle (1988) for interesting collections of effective data analyses carried out by these methods.

However there is a price to be paid for the great flexibility of nonparametric methods, which is that the smoothing parameter must be chosen. It is easy to see in Figure 1 which window width is appropriate, because that is a simulation example, where the underlying curve is available, but for real data sets, when one has little idea of what the underlying curve is like, this issue clearly becomes more difficult.

Most effective data analysis using nonparametric curve estimators has been done by choosing the smoothing parameter by a trial and error approach consisting of looking at several different plots representing different amounts of smoothness. While this approach certainly allows one to learn a great deal about the set of data, it can never be used to convince a skeptic in the sense that a hypothesis test can. Hence there has been a search for methods which use the data in some objective, or "automatic" way to choose the smoothing parameter.

This paper is a survey of currently available automatic smoothing

parameter selection techniques. There are many settings in which smoothing type estimators have been proposed and studied. Attention will be focussed here on the two most widely studied, which are density and regression estimation, because the lessons seem to be about the same for all settings. These problems are formulated mathematically in Section 2.

There are also many different types of estimators which have been proposed in each setting, see for example Prakasa Rao (1983), Silverman (1986), and Härdle (1988). However all of these have the property that, as with the moving average estimator in Figure 1, their performance is crucially dependent on choice of a smoothing parameter. Here again the lessons seem to be about the same, so focus is put on just one type of estimator, that is kernel based methods. These estimators are chosen because they are simple, intuitively appealing, and best understood. The form of these are given in Section 2. Other important estimators, to which the ideas presented here also apply include histograms, the various types of splines, and those based on orthogonal series.

To find out more about the intuitive and data analytic aspects of nonparametric curve estimation, see Silverman (1986) for density estimation and Härdle (1988) for regression. For an access to the rather large theoretical literature, the monograph by Prakasa Rao (1983) is recommended. Other monographs on curve estimation, some of which focus on some rather specialized topics, include Tapia and Thompson (1978), Wertz (1978), Devroye and Györfi (1984), Nadaraya (1983) and Devroye (1987). Survey papers have been written on density estimation

by Wegman (1972), Tarter and Kronmal (1976), Fryer (1977), Wertz and Schneider (1978) and Bean and Tsokos (1980). Collomb (1982, 1985) provides a survey of nonparametric regression.

Section 2 of this paper introduces notation. Section 3 discusses various possibilities for "the right amount of smoothing", and states an important asymptotic quantification of the smoothing problem. Section 4 introduces and discusses various methods for automatic bandwidth selection in the density estimation context. This is done for regression in Section 5. Section 6 discusses some hybrid methods and related topics.

## 2. Mathematical Formulation and Notation

The density estimation problem is mathematically formulated as follows. Use independent identically distributed observations,  $X_1, \dots, X_n$ , from a probability density  $f(x)$ , to estimate  $f(x)$ . The kernel estimator of  $f$ , as proposed by Rosenblatt (1956) and Parzen (1962), is given by

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^n K_h(x - X_i),$$

where  $K$  is often taken to be a symmetric probability density, and  $K_h(\cdot) = K(\cdot/h)/h$ . See Chapter 3 of Silverman (1986) for a good discussion of the intuition behind this estimator and its properties. The smoothing parameter in this estimator is  $h$ , often called the bandwidth or window width. Note that the estimator could have been defined without  $h$  appearing as a separate parameter, however because

the amount of smoothing is so crucial it is usually represented in this form.

One way of formulating the nonparametric regression problem is to think of using

$$Y_i = m(x_i) + \epsilon_i, \quad i = 1, \dots, n,$$

where the  $\epsilon_i$  are mean zero errors, to estimate the regression curve,  $m(x)$ . This setup is usually called "fixed design" regression. A widely studied alternative is "stochastic design" regression, in which the  $x_i$ 's are treated as random variables. While mathematical analysis of the two settings requires different techniques, the smoothing aspects tend to correspond very closely, so only the fixed design is explicitly formulated here. See Chapter 2 of Härdle (1988) for a formulation of the stochastic design regression problem. Kernel estimators in regression were introduced by Nadaraya (1964) and Watson (1964). One way to formulate them is as a weighted average of the form

$$\hat{m}_h(x) = \sum_{i=1}^n W_i(x, h) Y_i,$$

where the weights are defined by

$$W_i(x, h) = K_h(x - X_i) / \sum_{i=1}^n K_h(x - X_i).$$

See Section 3.1 of Härdle (1988) for discussion of this estimator, and a number of other ways of formulating a kernel regression estimator.

In both density and regression estimation, the choice of the kernel function,  $K$ , is of essentially negligible concern, compared to choice of the bandwidth  $h$ . This can be seen at intuitive level by again considering Figure 1. Note that if the shape of the weight functions

appearing at the bottom is changed, the effect on the estimator will be far less than is caused by a change in window width. See Section 3.3.2 of Silverman (1986) and section 4.5 of Härdle (1988) for a mathematical quantification and further discussion of this.

### 3. "Right" Answers

The traditional method of assessing the performance of estimators which use an automatic smoothing parameter selector, is to consider some sort of error criterion. The usual criteria may be separated into two classes, global and pointwise. As most applications of curve estimation call for a picture of an entire curve, instead of its value at one particular point, only global measures will be discussed here.

The most commonly considered global error criteria in density estimation are the Integrated Squared Error (i.e. the  $L^2$  norm),

$$\text{ISE}(h) = \int [\hat{f}_h - f]^2,$$

and its expected value, the Mean Integrated Squared Error

$$\text{MISE} = E(\text{ISE}(h)).$$

Related criteria are the Integrated Absolute Error (i.e. the  $L^1$  norm),

$$\text{IAE} = \int |\hat{f}_h - f|,$$

and its expected value,

$$\text{MIAE} = E(\text{IAE}(h)).$$

There are other possibilities such as weighted versions of the above, as well as the supremum norm, Hellinger distance, and the Kullback-Leibler distance.

In regression, one can study the obvious regression analog of the

above norms, and in addition there are other possibilities, such as the Average Squared Error,

$$ASE(h) = n^{-1} \sum_{j=1}^n [\hat{m}_h(x_j) - m(x_j)]^2,$$

and its expected value,

$$MASE(h) = E(ASE(h)).$$

For the rest of this paper, the minimizers of these criteria will be denoted by an  $h$  with the an appropriate subscript, e.g.  $h_{MISE}$ .

An important question is how much difference is there between these various criteria. In Section 6 of their Chapter 5, Devroye and Györfi (1984) report that there can be a very substantial difference between  $h_{MISE}$  and  $h_{MIAE}$  in density estimation. However, this point is not really settled as Hall and Wand (1988) feel that the difference between the bandwidths which minimize pointwise absolute and squared error are very close to being the same.

Given that there is an important difference between the squared error and the absolute error type criteria, there is no consensus on which should be taken as "the right answer". Devroye and Györfi (1984) point out a number of reasons for studying density estimation with absolute error methods. This has not gained wide acceptance though, one reason being that squared error criteria are much easier to work with from a technical point of view. The result of this is that all of the real theoretical breakthroughs in density estimation have come first from considering squared error criteria, then with much more work, the idea is extended to the absolute error case.

The issue of the difference between the random criteria, such as ISE and IAE, and their expected values, such as MISE and MIAE, seems more clear. In particular it has been shown by Hall and Marron (1987a) that  $h_{ISE}$  and  $h_{MISE}$  do converge to each other asymptotically, but at a very slow rate, and may typically be expected to be quite far apart. Here again there is no consensus about which should be taken as "the right answer". ISE has the compelling advantage that minimizing it gives the smoothing parameter which is best, for the set of data at hand, as opposed to being best only with respect to the average over all possible data sets, as with MISE. However, acceptance of ISE as the "right answer" is controversial because ISE is random, and two different experimenters, whose data have the same underlying distributions, will have two different "right answers". This type of reason is why statistical decision theory (see for example Ferguson 1967) is based on "risk" instead of on "loss". See Scott (1988) and Carter, Eagleson and Smith (1986) for further discussion of this issue.

One advantage of MISE is that it allows a very clean asymptotic summary of the smoothing problem. In particular, for kernel density estimation, if  $K$  is a probability density and  $f$  has two continuous derivatives, then as  $n \rightarrow \infty$  and  $h \rightarrow 0$ , with  $nh \rightarrow \infty$ ,

$$MISE(h) = AMISE(h) + o(AMISE(h)),$$

where

$$AMISE(h) = n^{-1}h^{-1}(\int K^2) + h^4(\int x^2 K)^2(\int (f'')^2)/4,$$

see for example (3.20) of Silverman (1986). Recall from Figure 1, that too small a window width results in too much sample variability. Note

that this is reflected by the first term (usually called the variance term) in AMISE becoming too large. On the other side, the fact that a too large window width gives too much bias, is reflected in the second term which gets large in that case.

There is a tendency to think of  $h_{AMISE}$  as being the same as  $h_{MISE}$ , and this seems to be usually nearly true, but can be quite far off sometimes. Scott (1986) has shown that for the lognormal density AMISE and MISE will still be very far apart even for sample sizes as large as a million.

#### 4. Density Estimation

In this section most of the automatic bandwidth selectors proposed for kernel density estimation are presented and discussed. It should be noted that most of these have obvious analogs for other types of estimators as well.

##### 4.1 Plug-in methods

The essential idea here is to work with  $AMISE(h)$  and plug in an estimate of the only unknown part, which is  $\int (f'')^2$ . Variations on this idea have been proposed and studied by Woodroffe (1970), Scott and Factor (1981), Krieger and Pickands (1981), and Sheather (1983, 1986). Most of the above authors consider the case of pointwise density estimation, but the essential ideas carry over to the global case.

A drawback to this approach, is that estimation of  $\int (f'')^2$  requires specification of a smoothing parameter. The argument is usually given

that the final estimator is less dependent on this secondary smoothing parameter, but this does not seem to have been very carefully investigated. An interesting approach to the problem is given in Sheather (1986).

A second weakness of the plug-in estimator is that it targets AMISE, which can be substantially different from MISE.

A major strength of the plug-in selector is that, if strong enough smoothness assumptions are made, then it seems to have much better sample variability properties than many of the selectors in the rest of section 4, see remark 4.6 of Hall and Marron (1987a).

For plug-in estimators in the absolute error setting, see Hall and Wand (1988) in the case of MAE, and Hall and Wand (1989) for MIAE.

#### 4.2 Psuedo Likelihood Cross-Validation

Also called Kullback-Leibler cross-validation, this was proposed independently by Habbema, Hermans and van den Broek(1974) and by Duin (1976). The essential idea is to choose that value of  $h$  which minimizes the psuedo-likelihood,

$$\prod_{j=1}^n \hat{f}_h(X_j).$$

However this has a trivial minimum at  $h = 0$ , so the cross-validation principle is invoked by replacing  $\hat{f}_h$  in each factor by the leave one out version,

$$\hat{f}_{h,j}(x) = (n-1)^{-1} \sum_{i \neq j}^n K_h(x-X_i).$$

Another viewpoint on why the leave-oneout estimator is appropriate here

is that the original criterion may be considered to be using the same observations to construct the estimator, as well as assess its performance. When the cross-validation principle (see Stone 1974) is used to attack this problem, we arrive at the modification based on using the leave-one-out estimator.

Schuster and Gregory (1981) have shown that this selector is severely affected by the tail behavior of  $f$ . Chow, Geman, and Wu (1983) demonstrated that if both the kernel and the density are compactly supported then the resulting density estimator will be consistent. The fact that this consistency can be very slow, and the selected bandwidth very poor, was demonstrated by Marron (1985), who proposed an efficient modification of the psuedo-likelihood based on some modifications studied by Hall (1982). Hall (1988a,b) has provided a nice characterization of the psuedo-likelihood type of cross-validation, by showing that it targets the bandwidth which minimizes the Kullback-Leibler distance between  $\hat{f}_h$  and  $f$ . Hall goes on to explore the properties of this bandwidth, and concludes that it may sometimes be appropriate for using a kernel estimate in the discrimination problem, but is usually not appropriate for curve estimation. For this reason, psuedo-likelihood currently seems to be of less current interest than the other smoothing parameter selectors considered here.

#### 4.3 Least Squares Cross-Validation

This was proposed independently by Rudemo (1982a) and by Bowman

(1984). The essential idea is to target

$$\text{ISE}(h) = \int \hat{f}_h^2 - 2 \int \hat{f}_h f + \int f^2.$$

The first term of this expansion is available to the experimenter, and the last term is independent of  $h$ . Using a method of moments estimate of the second term results in the criterion

$$\int \hat{f}_h^2 - 2 \sum_{j=1}^n \hat{f}_{h,j}(X_j),$$

which is then minimized to give a cross-validated smoothing parameter.

The fact that the bandwidth chosen in this fashion is asymptotically correct, under various assumptions for ISE, MISE, and AMISE has been demonstrated under various assumptions by Hall (1983, 1985), Stone (1984), Burman (1985) and Nolan and Pollard (1987) (see Stone 1985 for the histogram analog of this). Marron and Padgett (1987) have established the analogous result in the case of randomly censored data. A comparison to an improved version of the Kullback-Leibler cross-validation was done by Marron (1987a).

The main strength of this bandwidth is that it is asymptotically correct under very weak smoothness assumptions on the underlying density. Stone (1984) uses assumptions so weak that there is no guarantee that  $\hat{f}_h$  will even be consistent, but the bandwidth is still doing as well as possible in the limit. This translates in a practical sense into a type of robustness. The plug-in selector is crucially dependent on AMISE being a good approximation of MISE, but least squares cross-validation still gives good asymptotic performance, even in situations where the  $\text{MISE} \approx \text{AMISE}$  approximation is very bad.

A drawback to least squares cross-validation is that the score function has a tendency towards having several local minima, with some spurious ones often quite far over on the side of undersmoothing. This does not seem to be only a small sample aberration, as Scott and Terrell (1987) noticed it in their simulation study even for very large samples. For this reason it is recommended that minimization be done by a grid search through a range of  $h$ 's, instead of by some sort of computationally more efficient step-wise minimization algorithm.

Another major weakness of the least squares cross-validated smoothing parameter is that it is usually subject to a great deal of sample variability, in the sense that for different data sets from the same distributions, it will typically give much different answers. This has been quantified asymptotically by Hall and Marron (1987a), who show that the relative rate of convergence of the cross-validated bandwidth to either of  $h_{ISE}$  or  $h_{MISE}$  is excruciatingly slow. It is interesting though that the noise level is of about the same order as the relative difference between  $h_{ISE}$  and  $h_{MISE}$ . This is the result referred to in Section 3, concerning the practical difference between random error criteria and their expected values.

While the noise level of the cross-validated bandwidth is very large it is rather heartening that the same level exists for the difference between the two candidates for "optimal", in the sense that the exponents in the algebraic rate of convergence are the same. This leads one to suspect that the rate of convergence calculated by Hall and Marron(1987a) is in fact the best possible. This was shown, in a

certain minimax sense, by Hall and Marron (1987b). To keep this in perspective though, note that the constant multiplier of the rate of convergence of the optimal bandwidths to each other is typically smaller than for cross-validation. Since the rates are so slow, it is these constants which are really important. See Marron (1987b) for further discussion.

Recall that an attractive feature of the plug-in bandwidth selectors was their sample stability. These selectors have a faster rate of convergence to  $h_{AMISE}$  than the rate at which  $h_{ISE}$  and  $h_{MISE}$  come together. This does not contradict the above minimax result because this faster rate requires much stronger smoothness assumptions. In settings of the type which drive the minimax result, the plug-in selectors will be subject to much more sample noise, and also  $h_{AMISE}$  will be a very poor approximation to  $h_{AMISE}$ .

A somewhat surprising fact about the least squares cross-validated bandwidth is that, although its goal is  $h_{ISE}$ , these two random variables are in fact negatively correlated! This means that for those data sets where  $h_{ISE}$  is smaller than usual, the cross-validated bandwidth tends to be bigger than usual, and vice versa. This phenomenon was first reported in Rudemo (1982a), and has been quantified theoretically by Hall and Marron (1987a). An intuitive explanation for it has also been provided by Rudemo, in terms of "clusterings" of the data. If the data set is such that there is more clustering than usual, note that  $h_{ISE}$  will be larger than usual, because the spurious structure needs to be smoothed away, while cross-validation will pick a smaller

bandwidth than usual because it sees the clustering as some fine structure that can only be resolved with a smaller window. On the other hand, if the data set has less clustering than usual, then  $h_{ISE}$  will be smaller than usual, so as to cut down on bias, while cross-validation sees no structure, and hence takes a bigger bandwidth. An interesting consequence of this negative correlation is that if one could find a stable "centerpoint", then (if ISE is accepted as the right answer) it would be tempting to use a bandwidth which is on the opposite side of this from  $h_{CV}$ .

A last drawback of least squares cross-validation is that it can be very expensive to compute, especially when the recommended grid search minimization algorithm is used. Two approaches to this problem are the Fast Fourier Transform approximation ideas described in section 3.5 of Silverman (1986), and the Average Shifted Histogram approximation ideas described in Section 5.3 of Scott and Terrell (1987).

#### 4.4 Biased Cross-Validation

This was proposed and studied by Scott and Terrell (1987). It is a hybrid combining aspects of both plug-in methods, and also least squares cross-validation. The essential idea is to minimize, by choice of  $h$ , the following estimate of  $AMISE(h)$ ,

$$n^{-1}h^{-1}(\int K^2) + h^4(\int x^2 K)^2(\int (\hat{f}_h'')^2)/4.$$

This differs from the plug-in because the same  $h$  that is being assessed by this score function is used in the estimate of  $\int (f'')^2$ .

Scott and Terrell(1987) show that the biased cross-validated

bandwidth has sample variability with the same rate of convergence as least squares cross-validation, but with a typically much smaller constant coefficient. This is crucial as the rates are so slow that it is essentially the constant coefficients that determine performance of the selectors. Scott and Terrell (1987) also demonstrate the superior performance of biased cross-validation in some settings by simulation results.

A drawback of biased cross-validation is that, like the plug-in, its effective performance requires much stronger smoothness assumptions than required for least squares cross-validation. It seems possible that in settings where biased cross-validation is better than least squares cross-validation, the plug-in will be better yet, and in settings where the plug-in is inferior to least squares cross-validation, biased cross-validation will be as well, although this has not been investigated yet.

Another weak point of biased cross-validation is that for very small samples, on the order of  $n = 25$ , Scott and Terrell (1987) report that the method may sometimes fail all together, in the sense that there is no minimum. This seems to be not as bad as the spurious local minima that occur for least squares cross-validation, because at least it is immediately clear that something funny is going on. Also unlike the spurious minima in least squares cross-validation, this problem seems to disappear rapidly with increasing sample size.

#### 4.5 Oversmoothing

This idea has been proposed in Chapter 5, Section 6 of Devroye and Györfi (1984) and by Terrell and Scott (1985). The essential idea is to note that

$$h_{\text{AMISE}} = \left( \frac{\int k^2}{(\int x^2 k)^2 \int (f'')^2} \right)^{1/5} n^{-1/5}.$$

If the scale of the  $f$  distribution is controlled, say by rescaling so that its variance is equal to one, then  $\int (f'')^2$  has a lower bound over the set of all probability densities. When this lower bound is substituted, and the scale is taken properly into account, say using some estimate of the sample variance, then a bandwidth is arrived at which will be asymptotically bigger than any of the squared error notions of "optimal" described above. The version described here is that of Terrell and Scott, the Devroye and Györfi version is the  $L^1$  analog of this idea.

Terrell and Scott (1985) show that for unimodal densities, such as the Gaussian, the difference between the oversmoothed bandwidth and  $h_{\text{AMISE}}$  is often surprisingly small. Another benefit of this is that it is very stable across samples because the only place the data even enter are through the scale estimate, which has a fast parametric rate of convergence.

Of course the oversmoothed bandwidth has the obvious drawback that it can be very inappropriate for multimodal data sets, which unfortunately are the ones that are most interesting when data is being analyzed by density estimation techniques.

A possible application of the oversmoothed bandwidth, that has been suggested in some oral presentations given by David Scott, is that it can be used to provide an upper bound to the range of bandwidths considered by minimization based techniques such as the various types of cross-validation. The usefulness of this idea has not yet been investigated.

## 5 Regression Estimation

Note that two of the ideas proposed above for density estimation, the plug-in selectors and biased cross-validation, can be directly carried over to the regression setting. Neither of these ideas has been investigated yet (although see Müller and Stadtmüller 1987 for a local criterion based plug-in selector). In this section most of the automatic bandwidth selectors that have been considered for kernel regression estimation are presented and discussed. Again note that the ideas here have obvious analogs for other types of estimators as well.

### 5.1 Cross-validation

This was first considered in the nonparametric curve estimation context by Clark (1975) for kernel estimators and by Wahba and Wold (1975) for spline estimators. The essential idea is to use the fact that the regression function,  $m(x)$  is the best mean square predictor of a new observation taking at  $x$ . This suggests choosing the bandwidth which makes  $\hat{m}_h(x)$  a good predictor, or in other words taking the minimizer of the estimated prediction error,

$$\text{EPE}(h) = n^{-1} \sum_{j=1}^n [Y_j - \hat{m}_h(x_j)]^2.$$

This criterion has the same problem that was observed in section 4.2, namely that it has a trivial minimum at  $h = 0$ . As above this can be viewed as being caused by using the same data to both construct and assess the estimator, and a reasonable approach to this problem is provided by the cross-validation principle. Hence the cross-validated bandwidth is the one which minimizes the criterion obtained by replacing  $\hat{m}_h(x_j)$  by the obvious leave-one-out version. See Härdle and Marron (1985a) for a motivation of this criterion that is very similar in spirit to that for density estimation least squares cross-validation, as described in Section 4.3.

The fact that the bandwidth chosen in this fashion is asymptotically correct was established by Rice (1984) in the fixed design context, and by Härdle and Marron (1985a) in the stochastic design setting.

Härdle, Hall and Marron (1988) have shown that this method of bandwidth selection suffers from a large amount of sample variability, in a sense very similar to that described for the least squares cross-validated density estimation bandwidth in Section 4.3. In particular, the excruciatingly slow rate of convergence, and the negative correlation between the cross-validated and ASE optimal bandwidths are here also. See Marron (1987) for further discussion.

One thing that deserves further comment is that Rudemo has provided an intuitive explanation of the cause of the negative correlation in

this setting, which is closely related to his intuition for density estimation, as described near the end of Section 4.3. This time focus on the lag-one serial correlation of the actual residuals (i.e. on  $\rho(\epsilon_i, \epsilon_{i+1})$ ). Under the assumptions of independent errors, this will be zero on the average, but for any particular data set, the empirical value will typically be either positive or negative. Note that for data sets where the empirical serial correlation is positive, there will be a tendency for residuals to be positive in "clumps" (corresponding to the "clusters" described in Section 4.3). This clumping will require  $h_{ASE}$  to be larger than usual, so as to smooth them away. Another effect is that cross-validation will feel there is some fine structure present, which can only be recovered by a smaller bandwidth. For those data sets with a negative empirical correlation, the residuals will tend to alternate between positive and negative. The effect of this is that the sample variability will be smaller than usual, so ASE can achieve a smaller value by taking a small bandwidth which eliminates bias. On the other hand, cross-validation does not sense any real structure, and hence selects a relatively large bandwidth.

## 5.2 Model Selection Methods

There has been a great deal of attention to a problem very closely related to nonparametric smoothing parameter selection, which is often called "model selection". The basic problem can perhaps be best understood in the context of choosing the degree of a polynomial, for a least squares fit of a polynomial regression function, although the

largest literature concerns choice of the order of an ARMA fit in time series analysis. To see that these problems are very similar to bandwidth selection, note that when too many terms are entered into a polynomial regression, the resulting curve will be too wiggly, much as for the small bandwidth curve in Figure 1. On the other hand if too few terms are used, there will not be enough flexibility to recover all the features of the underlying curve, resulting in an estimate rather similar to the large bandwidth curve in Figure 1.

Given the close relationship between these problems, it is not surprising that there has been substantial cross-over between these two areas. The main benefit for nonparametric curve estimation has been motivation for a number of different bandwidth selectors. Rice (1984) has shown that the bandwidth selectors motivated by a number of these (based on the work of Akaike, Shibata and others), as well as the Generalized Cross Validation idea of Craven and Wahba (1979), all have the following structure. Choose the minimizer of a criterion of the form

$$\text{EPE}(h)\Psi(h),$$

where  $\text{EPE}(h)$  was defined in Section 5.1 above. The function  $\Psi(h)$  can be thought of as a correction factor, which has an effect similar to replacing the estimator by its leave-one-out version, as done by cross-validation, as described in Section 5.1 above. Rice (1984) shows that all of these bandwidths asymptotically come together, and also converge to  $h_{\text{MASE}}$ , see Li (1987) for a related result. Härdle and Marron (1985b) show that extra care needs to be taken with these

selectors when the design is not equally spaced, and the errors are not heteroscedastic. See Silverman (1985) for related discussion in the context of spline estimation.

Deeper properties, concerning the sample variability of these bandwidths, have been investigated by Härdle, Hall and Marron (1988). It is shown there that, in a much deeper sense than that of Rice (1984), all of these bandwidths are asymptotically equivalent to each other and also to the cross-validated bandwidth discussed in Section 5.1. Hence all properties described for the cross-validated bandwidth apply to these as well.

### 5.3 Unbiased Risk Estimation

This consists of yet another method of adjusting the estimated prediction error defined in section 5.1. This essential idea was proposed by Mallows (1973) in the model selection context, and comes from considering the expected value of  $EPE(h)$ . When this is done, it becomes apparent that the bias towards  $h$  too small can be corrected by minimizing a criterion of the form

$$EPE(h) + 2\hat{\sigma}^2 K(0)/nh,$$

where  $\hat{\sigma}^2$  is some estimate of the residual variance,  $\sigma^2 = E[\epsilon_i^2]$ .

There has been substantial work done on the nonparametric estimation of  $\sigma^2$ , see for example Kendall (1976). For more recent references, see Rudemo (1982b), Rice (1984), Gasser, Sroka and Jennen (1986) and Eagleson and Buckley (1987) for discussion of possible estimates of  $\sigma^2$ .

The results of Rice (1984) and Härdle, Hall and Marron (1988).

described above, apply here to this type of selector as well. For an interesting connection to Stein shrinkage estimation see Li(1985).

The main drawback to this type of selector is that it depends on an estimate of  $\sigma^2$ , although this should not be too troubling, because there are estimates available with a fast parametric convergence rate, so the amount of noise in this estimation is at least asymptotically negligible. Härdle, Hall and Marron (1988) have demonstrated that there is a sense in which the other bandwidth selectors are essentially doing the same thing as the unbiased risk estimator, except that the variance estimation is essentially being provided by  $EPE(h)$ .

An advantage of this selector over the selectors in sections 5.1 and 5.2 is that it appears to handle settings in which reasonable values of the bandwidth are close to  $h = 0$ . This happens typically when there is very small error variance, so not much local averaging needs to be done. It is immediately clear that cross-validation suffers in this type of context, but it can also be seen that the other selectors have similar problems. These issues have not been well investigated yet.

## 6. Extensions, Hybrids and Hopes for the Future

There are many possibilities for modifying the above selectors, in the hopes of improving them.

Bhattacharya and Mack (1987) have studied a stochastic process that is closely related to the plug-in bandwidth selectors (they work explicitly with the nearest neighbor density estimator, but it appears that analogous ideas should hold for conventional kernel density and

regression estimators as well). This gives a certain linear model, which is then used to give an improved bandwidth. It seems there may be room for improvement of this type of some of the other bandwidth selectors described above as well.

Burman (1988) provides a detailed analysis and suggests and improvement of v-fold cross-validation. The basic idea here is to replace the leave-one-out estimators by leave-several-out versions, and then assess the performance against each of the left out observations. This has the distinct advantage that it cuts down dramatically on the amount of computation required, at least if direct implementation is used (there seems to be less gain if an algorithm of one of the types described at the end of Section 4.3 is used). If too many observations are left out, then note that the selected smoothing parameter should be reasonably good, except that it will be appropriate for the wrong sample size (the size of the sample minus those left out). Burman provides a nice quantification of this, and goes on to use the quantification to provide an appropriate correction factor. Burman also provides similar results for another modification of cross-validation based on "repeated learning testing" methods.

Another means of modifying cross-validation, and also a number of the other automatic bandwidth selectors, is through partitioning, as proposed by Marron (1987c). The basic idea here is to first partition the data into subsets. In density estimation this could be done randomly, while it may be more appropriate to take every k-th point in regression. Next construct the cross-validation score for each sample

separately, and find the minimizer of the average of these score functions. When this is rescaled to adjust for the fact that the subset cross-validation scores are for much smaller sample sizes, the resulting bandwidth will often have, up to a point, much better sample stability than ordinary cross-validation. A drawback to this approach is that one must decide on the number of subsets to use, and this problem seems to closely parallel that of smoothing parameter selection.

N. I. Fisher has pointed out that partitioned cross-validation seems subject to a certain inefficiency, caused by the fact that observations are not allowed to "interact" with observations in the other subsets. He then proposed overcoming this problem by replacing the average over cross-validation scores for the partition subsets by an average of the scores over all possible subsets of a given size. Of course there are far too many subsets to actually calculate this score function, so an average over some randomly selected subsets should probably be implemented. This idea has yet to be analyzed, although again the subsample size will clearly be an important issue.

Wolfgang Härdle has proposed another possibility along these lines, for density estimation. The idea is to minimize a cross-validation score based on a subsample consisting of say every  $k$ -th order statistic, or perhaps of averages of blocks of  $k$  order statistics. The resulting sample would be much more stable with respect to the type of clusterings which drive Rudemo's intuition, regarding the noise in least-squares cross-validation described near the end of Section 4.3. This idea has also not been investigated, and here again choice of  $k$  seems crucial.

## 7. Conclusions

It should be clear from the above that the field of smoothing parameter selection is still in its infancy. While many methods have been proposed, none has emerged as clearly superior. There is still much comparison to be done, and many other possibilities to be investigated.

The implications, in terms of actual data analysis, of what is known currently about automatic methods of smoothing parameter selection, are that there is still no sure-fire replacement for the traditional trial and error method. This is where one plots several smooths, and then chooses one based on personal experience and opinion. Indeed none of these automatic methods should be used without some sort of verification of this type.

Scott and Terrell (1987) have suggested the reasonable idea of looking at several automatically selected bandwidths, with the idea that they are probably acceptable when they agree, and there is at least a good indication that special care needs to be taken when they disagree. There are many things yet to be investigated in connection to this idea, especially in terms of how correlated all these automatically selected bandwidths are to each other. Also there is the issue of how many different bandwidths the user is willing to consider. Hopefully the field can at least be narrowed somewhat.

## 8. References

- Bean, S. J. and Tsokos, C. P. (1980), "Developments in nonparametric density estimation," *International Statistical Review*, 48, 267-287.
- Bhattacharya, P. K. and Mack, K. P. (1987), "Weak convergence of k-NN density and regression estimators with varying k and applications," *Annals of Statistics*, 15, 976-994.
- Bowman, A. (1984), "An alternative method of cross-validation for the smoothing of density estimates," *Biometrika*, 65, 521-528.
- Burman, P. (1985), "A data dependent approach to density estimation," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 69, 609-628.
- Burman, P. (1988), "Estimation of the optimal transformations using v-fold cross-validation and repeated learning testing methods," unpublished manuscript.
- Chow, Y. S., Geman, S. and Wu, L. D. (1983), "Consistent cross-validated density estimation," *Annals of Statistics*, 11, 25-38.
- Clark, R. M. (1975), "A calibration curve for radio carbon dates," *Antiquity*, 49, 251-266.
- Collomb, G. (1981), "Estimation non parametrique de la regression: revue," *International Statistical Review*, 49, 75-93.
- Collomb, G. (1985), "Nonparametric regression: an up-to-date bibliography," *Statistics*, 16, 309-324.
- Craven, P. and Wahba, G. (1979), "Smoothing noisy data with spline functions," *Numerische Mathematik*, 31, 377-403.
- Devroye, L. and Györfi, L. (1984). *Nonparametric Density Estimation: The  $L_1$  View*. Wiley, New York.
- Devroye, L. (1987), *A course in density estimation*, Birkhauser, Boston.
- Duin, R. P. W. (1976), "On the choice of smoothing parameters of Parzen estimators of probability density functions," *IEEE Transactions on Computers*, C-25, 1175-1179.
- Eagleson, G. K. and Buckley, M. J. (1987), "Estimating the variance in nonparametric regression," unpublished manuscript.

- Ferguson, T. S. (1967), *Mathematical Statistics, a Decision Theoretic Approach*, Academic Press, New York.
- Fryer, M. J. (1977). "A review of some non-parametric methods of density estimation," *Journal of the Institute of Mathematics and its Applications*, 20, 335-354.
- Gasser, T., Sroka, L. and Jennen, C. (1986), "Residuals variance and residual pattern in nonlinear regression," *Biometrika*, 73, 625-633.
- Habbema, J. D. F., Hermans, J. and van den Broek, K. (1984), "A stepwise discrimination analysis program using density estimation," *Compstat 1974: Proceedings in Computational Statistics*, 101-110, Physica Verlag, Vienna.
- Härdle, W. (1988), *Applied nonparametric regression*.
- Härdle, W. and Marron, J. S. (1985a), "Optimal bandwidth selection in nonparametric regression function estimation," *Annals of Statistics*, 12, 1465-1481.
- Härdle, W. and Marron, J. S. (1985b), "Asymptotic nonequivalence of some bandwidth selectors in nonparametric regression," *Biometrika*, 72, 481-484.
- Härdle, W., Hall, P. and Marron, J. S. (1988), "How far are automatically chosen regression smoothers from their optimum?," to appear with discussion, *Journal of the American Statistical Association*.
- Hall, P. (1982), "Cross-validation in density estimation," *Biometrika*, 69, 383-390.
- Hall, P. (1983), "Large sample optimality of least square cross-validation in density estimation," *Annals of Statistics* 11, 1156-1174.
- Hall, P. (1985), "Asymptotic theory of minimum integrated square error for multivariate density estimation," *Proceedings of the Sixth International Symposium on Multivariate Analysis at Pittsburgh*, 25-29.
- Hall, P. (1988a), "On the estimation of probability densities using compactly supported kernels," unpublished manuscript.
- Hall, P. (1988b), "On Kullback-Leibler loss and density estimation," unpublished manuscript.
- Hall, P. and Marron, J. S. (1987a), "Extent to which least-squares cross-validation minimises integrated square error in nonparametric

- density estimation," *Probability Theory and Related Fields*, 74, 567-581.
- Hall, P. and Marron, J. S. (1987b), "On the amount of noise inherent in bandwidth selection for a kernel density estimator," *Annals of Statistics*, 15, 163-181.
- Hall, P. and Marron, J. S. (1987c), "Estimation of integrated squared density derivatives", *Statistics and Probability Letters*, 6, 109-115.
- Hall, P. and Wand, M. (1988), "On the minimization of absolute distance in kernel density estimation," to appear in *Statistics and Probability Letters*.
- Hall, P. and Wand, M. (1989), "Minimizing  $L_1$  distance in nonparametric density estimation," to appear in *Journal of Multivariate Analysis*.
- Kendall, M. S. (1976), *Time Series*, Griffin, London.
- Krieger, A. M. and Pickands, J. (1981), "Weak convergence and efficient density estimation at a point," *Annals of Statistics*, 9, 1066-1078.
- Li, K. C. and Hwang, J. (1984). "The data smoothing aspects of Stein estimates," *Annals of Statistics*, 12, 887-897.
- Li, K. C. (1985). "From Stein's unbiased risk estimates to the method of generalized cross-validation," *Annals of Statistics*, 13, 1352-1377.
- Li, K. C. (1987). "Asymptotic optimality for  $C_p$ ,  $C_L$ , cross-validation and generalized cross-validation: discrete index set," *Annals of Statistics*, 15, 958-975.
- Mallows, C. L. (1973). "Some comments on  $C_p$ ," *Technometrics*, 15, 661-675.
- Marron, J. S. (1985). "An asymptotically efficient solution to the bandwidth problem of kernel density estimation," *Annals of Statistics*, 13, 1011-1023.
- Marron, J. S. (1986). "Will the art of smoothing ever become a science?", *Function estimates*, (J. S. Marron, ed.) American Mathematical Society Series: Contemporary Mathematics, 9, 169-178.
- Marron, J. S. (1987a). "A comparison of cross-validation techniques in density estimation", *Annals of Statistics*, 15, 152-162.

- Marron, J. S. (1987b), "What does optimal bandwidth selection mean for nonparametric regression estimation?", *Statistical data analysis based on the  $L^1$  norm and related methods*, (Y. Dodge, ed.) North Holland, Amsterdam.
- Marron, J. S. (1987c), "Partitioned cross-validation", North Carolina Institute of Statistics, Mimeo Series # 1721.
- Marron, J. S. and Padgett, W. J. (1987), "Asymptotically optimal bandwidth selection for kernel density estimators from randomly right-censored samples," *Annals of Statistics*, 15, 1520-1535.
- Müller, H. G. and Stadtmüller, U. (1985), "Variable bandwidth kernel estimators of regression curves," unpublished manuscript.
- Nadaraya, E. A. (1964), "On estimating regression," *Theory of Probability and its Application*, 9, 141-142.
- Nadarya, E. A. (1983), *Neparametrischeskoe otsenivanie plotnosti veroyanostei i krivoi regressii*, Izdatelstvo Tbilisskogo universiteta: Tbilisi.
- Nolan, D. and Pollard, D. (1987), "U-processes: rates of convergence," *Annals of Statistics*, 15, 780-799.
- Parzen, E. (1962), "On estimation of a probability density function and mode," *Annals of Mathematical Statistics*, 33, 1065-1076.
- Prakasa Rao, B. L. S. (1983), *Nonparametric Functional Estimation*, Academic Press, New York.
- Rice, J. (1984), "Bandwidth choice for nonparametric regression," *Annals of Statistics*, 12, 1215-1230.
- Rosenblatt, M. (1956), "Remarks on some non-parametric estimates of a density function," *Annals of Mathematical Statistics*, 27, 832-837.
- Rosenblatt, M. (1971), "Curve estimates," *Annals of Mathematical Statistics*, 42, 1815-1842.
- Rudemo, M. (1982a), "Empirical choice of histograms and kernel density estimators," *Scandinavian Journal of Statistics*, 9, 65-78.
- Rudemo, M. (1982b), "Consistent choice of linear smoothing methods," Report 82-1, Department of Mathematics, Royal Danish Agricultural and Veterinary University, Copenhagen.
- Schuster, E. A. and Gregory, C. G. (1981), "On the nonconsistency of

- maximum likelihood nonparametric density estimators," *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface* (W. F. Eddy ed.) Springer Verlag, New York, 295-298.
- Scott, D. W. (1986), "Handouts for ASA short course in density estimation," Rice University Technical Report 776-331-86-2.
- Scott, D. W. (1988), discussion of "How far are automatically chosen regression smoothers from their optimum?," by Härdle, W., Hall, P. and Marron, J. S. to appear *Journal of the American Statistical Association*.
- Scott, D. W. and Factor, L. E. (1981), "Monte Carlo study of three data-based nonparametric probability density estimators," *Journal of the American Statistical Association*, 76, 9-15.
- Scott, D. W. and Terrell, G. R. (1987). "Biased and unbiased cross-validation in density estimation," *Journal of the American Statistical Association*, 82, 1131-1146.
- Sheather, S. J. (1983), "A data-based algorithm for choosing the window width when estimating the density at a point," *Computational Statistics and Data Analysis*, 1, 229-238.
- Sheather, S. J. (1986). "An improved data-based algorithm for choosing the window width when estimating the density at a point," *Computational Statistics and Data Analysis*, 4, 61-65.
- Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York.
- Stone, C. J. (1984), "An asymptotically optimal window selection rule for kernel density estimates," *Annals of Statistics*, 12, 1285-1297.
- Stone, C. J. (1985), "An asymptotically optimal histogram selection rule," *Proceedings of the Berkeley Symposium in Honor of Jerzy Neyman and Jack Keifer*.
- Stone, M. (1974), "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society, Series B*, 36, 111-147.
- Tapia, R. A. and Thompson, J. R. (1978), *Nonparametric Probability Density Estimation*, The Johns Hopkins University Press, Baltimore.
- Tarter, M. E. and Kronmal, R. A. (1976), "An introduction to the implementation and theory of nonparametric density estimation," *The American Statistician*, 30, 105-112.
- Terrell, G. R. and Scott, D. W. (1985), "Oversmoothed density

estimates," *Journal of the American Statistical Association*, 80, 209-214.

Wahba, G. and Wold, S. (1975), "A completely automatic french curve: fitting spline functions by cross-validation," *Communications in Statistics*, 4, 1-17.

Watson, G. S. (1964), "Smooth Regression Analysis," *Sankhyā*, series A, 26, 359-372.

Wegman, E. J. (1972), "Nonparametric probability density estimation: I. a summary of the available methods," *Technometrics*, 14, 533-546.

Wertz, W. (1978), *Statistical Density Estimation: A Survey*, *Angewandte Statistique und Okonometrie* 13, Vandenhoeck und Ruprecht.

Wertz, W. and Schneider, B. (1979), "Statistical density estimation: a bibliography," *International Statistical Review*, 49, 75-93.

Woodroffe, M. (1970), "On choosing a delta sequence," *Annals of Mathematical Statistics*, 41, 1665-1671

Caption

Figure 1: Simulated regression setting. Solid curve is underlying regression. Residuals are Gaussian. Dashed curves are moving wieghted averages, with Gaussian weights, represented at the bottom. Standard deviations in the weight functions are: Figure 1b, 0.015; Figure 1c, 0.04; Figure 1d, 0.12.

FIGURE 1a

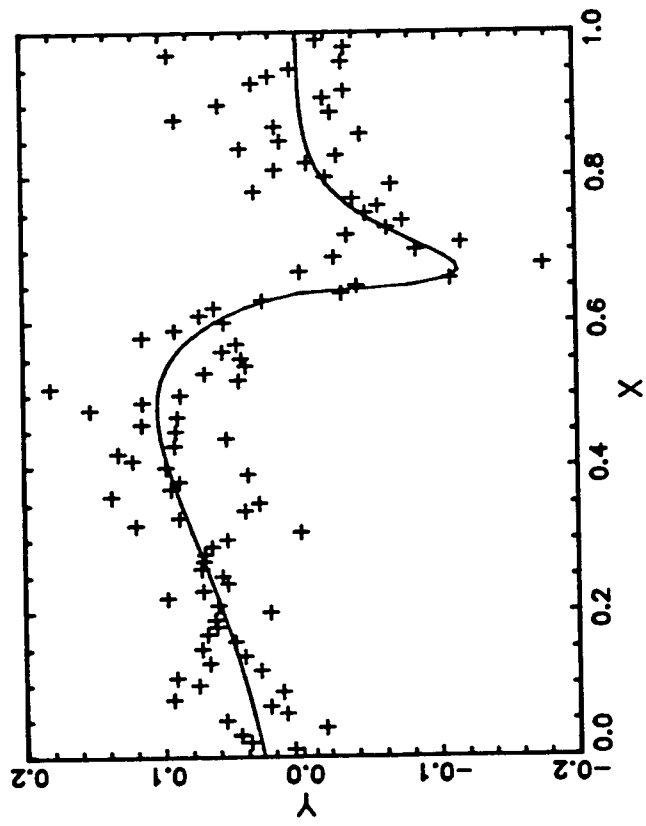


FIGURE 1b

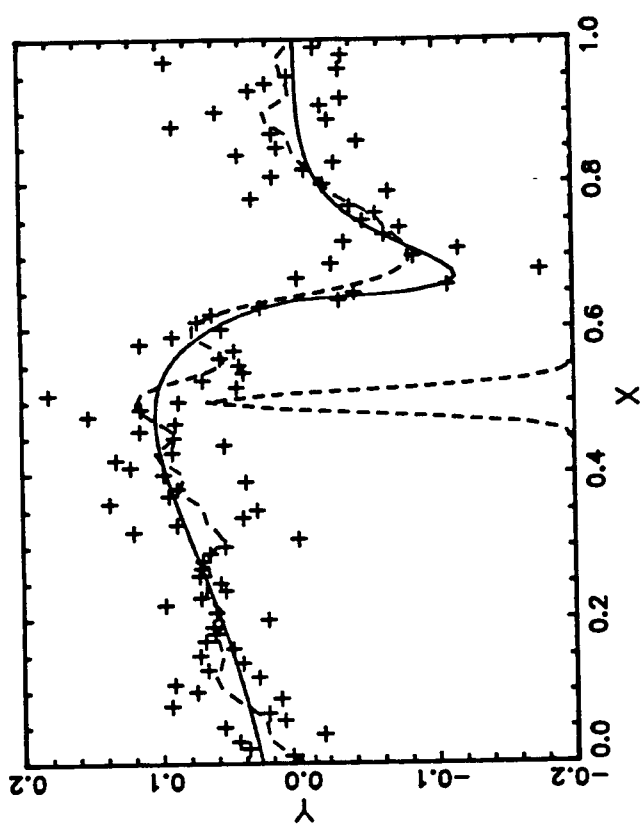


FIGURE 1c

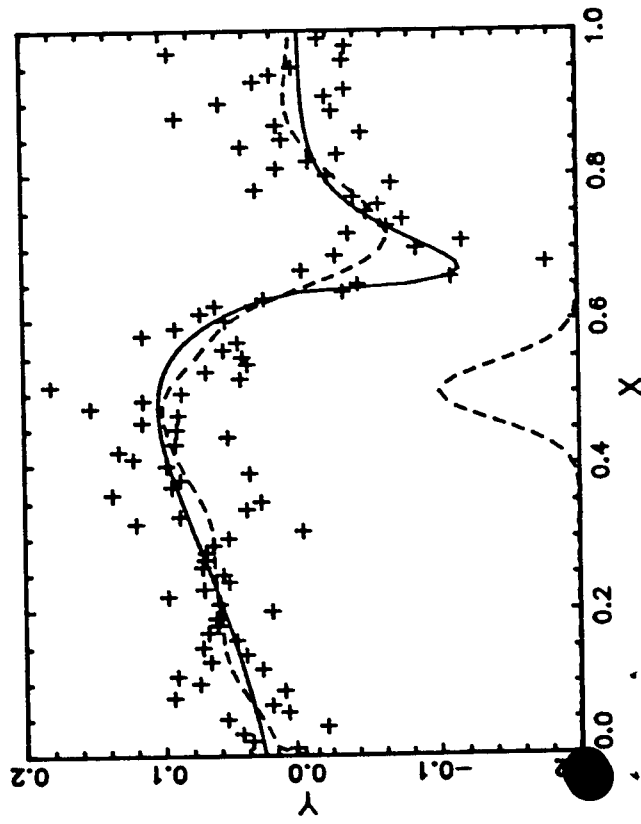


FIGURE 1d

