THE INSTITUTE
OF STATISTICS

THE CONSOLIDATED UNIVERSITY
OF NORTH CAROLINA

DATA-DRIVEN BANDWIDTH SELECTION IN LOCAL POLYNOMIAL FITTING: VARIABLE BANDWIDTH AND SPATIAL ADAPTATION

by

Jianqing Fan

and

Irene Gijbels

June 1993

Mimeo Series #2301

DEPARTMENT OF STATISTICS
Chapel Hill, North Carolina

Mimeo
Series
2301

Jianqing Fan & Irene
Gijbels

Data-Driven Bandwidth Selection in Local Polynomial Fitting Variable Bandwidth and Spatial Adaptation

Name	Date

Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation

Jianqing Fan[†]
Department of Statistics
University of North Carolina
Chapel Hill, N.C. 27599-3260

Irène Gijbels
Institut de Statistique
Université Catholique de Louvain
B-1348 Louvain-la-Neuve, Belgium

June 16, 1993

Abstract

When estimating a mean regression function and its derivatives, locally weighted least squares regression has been proven to be a very attractable technique. An important issue is how to determine the smoothing parameter or bandwidth. In case of estimating curves with a complicated structure, a variable bandwidth is desirable. Furthermore, the bandwidth should be 'instructed' by the data itself. Recent advancement of nonparametric smoothing techniques inspired us to propose such a data-driven bandwidth selection procedure, which can be used to select both constant and variable bandwidths. The idea is based on an Extended Cross-Validation criteria along with a natural approximation of the bias and variance of the estimator. The procedure itself can be applied to select bandwidths not only for estimating the regression curve, but also for estimating its derivatives. The resulting estimation procedure possesses the necessary flexibility for capturing complicated shapes of curves. This is illustrated via a large variety of testing examples. Those include highly spatial-variable examples where the variable bandwidth should be used. The results are also compared with wavelets techniques, and it seems that our results are at least comparable with those produced by the wavelets. In other words, local polynomial regression along with data-driven variable bandwidth has a similar spatial adaptation feature as wavelets.

Completed while the first author was visiting the Institute of Statistics, Louvain-la-Neuve, Belgium.

[†] Supported by NSF Grant DMS-9203135.

Abbreviated title: Variable bandwidth selection.

AMS 1991 subject classification. Primary 62G07. Secondary 62E25, 62H99.

Key words and phrases: Data-driven bandwidth selection, derivative estimation, Extended Cross-Validation, locally weighted regression.

1 Introduction

1.1 Objectives

In this paper the association between variables is exploited via describing the mean regression function and its derivatives. No preassumption about the form of this function is made – the complexity of the model will be determined completely by the data. We will use a particular nonparametric smoothing technique — local polynomial regression. The reasons for this choice of smoothing method are ample: nice minimax properties, no need for boundary modifications, applicable for various design-situations, easy to interpret, to implement, and to adapt to estimating derivatives. All nonparametric smoothing techniques involve the choice of a smoothing parameter or bandwidth. It is well-known that the choice of the smoothing parameter is rather crucial in the performance of the estimation procedure. Hence a very decisive question is how to choose this parameter.

The aim of this paper is to address this question when using local polynomial fitting for estimating the regression function and its derivatives. A bandwidth can be chosen to remain constant or to vary with the considered location point or with the data. Müller and Stadtmüller (1987) discussed the issue of local variable bandwidth for convolution type estimators for regression curves. Gasser, Kneip and Köhler (1991), Sheather and Jones (1991), Hall, Sheather, Jones and Marron (1991) and Brockmann et al. (1993) consider data-driven bandwidth selection rules based on "plug-in" techniques in a different setup. See also Vieu (1991) and Ruppert, Sheather and Wand (1993). For a survey on recent advancement of bandwidth selection see Jones, Marron and Sheather (1992) and references therein.

A constant bandwidth can be sufficient if the unknown curve is not to wiggly, i.e. has a high degree of smoothness. Such a bandwidth however fails to do a good job, when the unknown curve has a rather complicated structure. In order to capture the complexity of such a curve, a variable bandwidth is a necessity. This point will also be very clear from the examples we present at the end of this paper. Those include the examples discussed by Donoho and Johnstone (1992), which they used to illustrate the performance of their Wavelets-packages. Here, we analyse these examples using our proposed methodology based on local polynomial approximations. The reasons for presenting these examples are twofold. First of all, the theoretical curves are quite unsmooth or show many alterations, and are hence a good test for a newly proposed methodology. Secondly, it is interesting to compare the performance of both methods, wavelets and local polynomial fitting. It turns out that our results are at least comparable to wavelets techniques.

We will introduce a procedure which selects the, constant or variable (i.e. varying with the location point) bandwidth in a fully automatic way. The ideas for the developed procedure were inspired by the pioneering work on Generalized Cross-Validation by Wahba (1977) and Craven and Wahba (1979), and are related to those in Müller (1988). The procedure relies on the ideal assess of bias and variance discussed in Fan and Gijbels (1993). The proposed methodology is applicable when dealing with estimating the unknown regression function or any of its derivatives, as will be demonstrated. The method is based on a quantity called Extended Cross-Validation. The motivation and theoretical foundations for considering such a quantity rely on a thorough study of bias and variance (exact and approximated) of the estimators.

In organizing the paper, we opted for a presentation which highlights the main ideas leading to the proposed procedure. Details are left for secondary reading and are therefore collected in a last section. In the remainder of this section we give the notations involved with the local polynomial approximation method. The next section then introduces and motivates the Extended Cross-Validation quantity. Section 3 summarizes the ideal assessment of the bias and variance. The materials established in Sections 2 and 3 will serve as building blocks for the automatic bandwidth selection procedure described in Section 4. The performance of the proposed procedure is investigated extensively in Section 5. A large variety of testing examples is provided, which is meant to give the reader a clear and detailed picture of the strength of the methodology.

1.2 Local polynomial approximation

Let X and Y be two random variables whose relationship can be modeled as

$$Y = m(X) + \sigma(X)\varepsilon, \qquad E\varepsilon = 0 \quad \text{and} \quad \mathrm{Var}(\varepsilon) = 1,$$

where X and ε are independent. Of interest is to estimate the regression function $m(x) = E(Y|X = x)$ and its derivatives, based on $(X_1, Y_1), \cdots, (X_n, Y_n)$, a random sample from the population (X, Y). We use local polynomial fitting as estimation method, since it has various nice features (see e.g. Stone (1977), Fan (1992, 1993), Fan and Gijbels (1992), Ruppert and Wand (1992) and Fan et. al (1993)). The papers by Cleveland (1979) and Cleveland and Devlin (1988) contain a variety of nice examples showing the performance of locally-weighted regression in various fields of application. The particular class of locally-weighted running-line smoothers were discussed in Hastie and Tibshirani (1986).

If the $(p+1)^{th}$ derivative of $m(x)$ at the point x_0 exists, we approximate $m(x)$ locally by a polynomial of order p:

$$m(x) \approx m(x_0) + m'(x_0)(x - x_0) + \cdots + m^{(p)}(x_0)(x - x_0)^p/p!, \tag{1.1}$$

for x in a neighborhood of x_0. One then carries through a local polynomial regression

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} \beta_j (X_i - x_0)^j \right)^2 K\left(\frac{X_i - x_0}{h}\right), \tag{1.2}$$

where $K(\cdot)$ denotes a nonnegative weight function and h — a smoothing parameter — determines the size of the neighborhood of x_0. If $\{\hat{\beta}_\nu\}$ denotes the solution to the above weighted least squares problem, then it is clear from (1,1) that $\nu!\hat{\beta}_\nu$ estimates $m^{(\nu)}(x_0), \nu = 0, \cdots, p$.

It is more convenient to write the above least squares problem in matrix notation. Denote by $\mathbf{W}$ the diagonal matrix with entries $W_i \equiv K\left(\frac{X_i - x_0}{h}\right)$. Let $\mathbf{X}$ be the design matrix whose $(l, j)^{th}$ element is $(X_l - x_0)^{j-1}$ and put $\mathbf{y} = (Y_1, \cdots, Y_n)^T$. Then, the weighted least squares problem (1.2) can be written in matrix form as:

$$\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)^T \mathbf{W} (\mathbf{y} - \mathbf{X}\beta),$$

where $\beta = (\beta_0, \cdots, \beta_p)^T$. Ordinary least squares theory provides the solution

$$\hat{\beta} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}$$

whose conditional mean and variance are:

$$\left\{ \begin{array}{lcl} E(\hat{\beta}|X_1, \cdots, X_n) & = & (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{m} = \beta + (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{r} \\ \mathrm{Var}(\hat{\beta}|X_1, \cdots, X_n) & = & (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}^T\Sigma\mathbf{X})(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1} \end{array} \right. , \qquad (1.3)$$

where $\mathbf{m} = (m(X_1), \cdots, m(X_n))^T$, $\mathbf{r} = \mathbf{m} - \mathbf{X}\beta$, the residual of the local polynomial approximation, and $\Sigma = \mathrm{diag}\Big(K^2((X_i - x_0)/h)\sigma^2(X_i)\Big)$.

2 Extended Cross-Validation

We now introduce the Extended Cross-Validation quantity which will form one of the building blocks in the selection procedure.

Before introducing this quantity let us look at the theoretical optimal variable bandwidth which would be the ideal one to work with. The theoretical variable bandwidth for estimating $\beta_\nu = m^{(\nu)}(x_0)/\nu!$ is the one that minimizes the theoretical Mean Squared Error (MSE) which can be approximated by

$$\beta_{p+1}^2 b_\nu^2 h^{2(p+1-\nu)} + a_\nu \frac{\sigma^2(x_0)}{f_X(x_0)} \frac{1}{nh^{1+2\nu}}, \qquad (2.1)$$

where $f_X(\cdot)$ is the marginal density of X, i.e. the design density. Here we introduced the notation a_ν for the $(\nu+1)^{th}$ diagonal element of the matrix $S^{-1}S^*S^{-1}$, where S (respectively S^*) is a $(p+1) \times (p+1)$ matrix whose $(i,j)^{th}$ element is s_{i+j-2} (respectively ν_{i+j-2}), with $s_j = \int u^j K(u)du$ and $\nu_j = \int u^j K^2(u)du$. Further, b_ν is the $(\nu+1)^{th}$ element of the $(p+1)$-vector $S^{-1}(s_{p+1}, \cdots, s_{2p+1})^T$. See Ruppert and Wand (1992) and Fan et al. (1993).

This approximated MSE is minimized at

$$h_{\nu,\mathrm{opt}}(x_0) = \left(\frac{(2\nu+1)a_\nu\sigma^2(x_0)}{2(p+1-\nu)b_\nu^2\beta_{p+1}^2 n f_X(x_0)} \right)^{\frac{1}{2p+3}}. \qquad (2.2)$$

This theoretical optimal bandwidth does depend on unknown quantities. Plug-in methods rely on estimating these quantities first and then substituting them into the expression.

Our goal is now to come up with a statistic for which the minimizer leads to an estimator for the theoretical optimal bandwidth. Such a statistic is provided by the Extended Cross-Validation quantity which is based on the normalized weighted residual sum of squares:

$$\hat{\sigma}^2(x_0) = \frac{1}{\mathrm{tr}(\mathbf{W} - (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^2\mathbf{X})}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 K\left(\frac{X_i - x_0}{h}\right), \qquad (2.3)$$

with $\hat{\mathbf{y}} = (\hat{Y}_1, \cdots, \hat{Y}_n)^T = \mathbf{X}\hat{\beta}$. The Extended Cross-Validation (ECV) is defined as

$$\mathrm{ECV}(x_0; h) = \hat{\sigma}^2(x_0)\,(1 + (p+1)V_{n,0}), \qquad (2.4)$$

where $V_{n,\nu}$ is the $(\nu + 1)^{th}$ diagonal element of the matrix $S_n^{-1}S_n^*S_n^{-1}$, with $S_n = \mathbf{X}^T\mathbf{W}\mathbf{X}$ and $S_n^* = \mathbf{X}^T\mathbf{W}^2\mathbf{X}$.

The intuition behind statistic (2.4) is as follows. When the local polynomial does not fit well, i.e. the bandwidth h is too large, the bias is large and hence also the residual sum of squares $\hat{\sigma}^2(x_0)$. When the bandwidth h is too small, the variance $V_{n,0}$ tends to be larger. So the ECV-quantity does 'protect' for both extreme choices.

The theoretical justification for the quantity $\mathrm{ECV}(x_0; h)$ routes back to the following result, which will be proved in Section 6.

Theorem 1

Suppose that $\sigma^2(x) = \sigma^2(x_0)$ in a neighbourhood of x_0. If $h_n \to 0$, and $nh_n \to \infty$, then

$$E(ECV(x_0; h)|X_1, \cdots, X_n) = \sigma^2(x_0) + C_p\beta_{p+1}^2 h_n^{2p+2} + (p+1)a_0\frac{\sigma^2(x_0)}{nh_n f_X(x_0)} + o_P(h_n^{2p+2} + \frac{1}{nh_n}),$$

where

$$C_p = \frac{s_{2p+2} - (s_{p+1}, \cdots, s_{2p+1})S^{-1}(s_{p+1}, \cdots, s_{2p+1})^T}{s_0}. \qquad (2.5)$$

The minimizer of $\mathrm{E}(\mathrm{ECV}(x_0; h)|X_1, \cdots, X_n)$ is approximately equal to

$$h_o(x_0) = \left(\frac{a_0\sigma^2(x_0)}{2C_p\beta_{p+1}^2 n f_X(x_0)}\right)^{\frac{1}{2p+3}}.$$

Now, the relationship between $h_{\nu,\mathrm{opt}}(x_0)$ in (2.2) and $h_o(x_0)$ is very simple:

$$\begin{aligned} h_{\nu,\mathrm{opt}}(x_0) &= \left(\frac{(2\nu+1)}{(p+1-\nu)} \frac{a_\nu}{a_0} \frac{C_p}{b_\nu^2} \right)^{\frac{1}{2p+3}} h_o(x_0) \\ &\equiv \mathrm{adj}_{p,\nu} h_o(x_0). \end{aligned}$$

Here, $p-\nu$ must be odd. This is natural since the estimator with $p-\nu$ even is inadmissible.

Remark that the adjusting constants $\mathrm{adj}_{p,\nu}$, appearing in this expression depend only on the kernel function K, and hence can be calculated explicitly. As an illustration, we present in Table 1 below, these constants for the Epanechnikov and the Gaussian kernel. The above relationship and Theorem 1 form the core of the theoretical motivation for the ECV-quantity.

Table 1: Adjusting constants for Epanechnikov and Gaussian kernel

Epanechnikov Kernel							
p $p-\nu$	1	2	3	4	5	6	7
1	.8941	.7643	.7776	.7639	.7827	.7835	.7989
3			.8718	.8324	.8384	.8297	.8392
5					.8819	.8639	.8679
7							.8932
Gaussian Kernel							
1	1.000	.8403	.8285	.8085	.8146	.8098	.8159
3			.9554	.8975	.8846	.8671	.8652
5					.9495	.9165	.9055
7							.9470

The use of the ECV-quantity to estimate a global, i.e. constant, bandwidth becomes transparent now. Suppose we want to estimate $m^{(\nu)}(\cdot)$ on $[c,d]$. Then, find $\hat{h}$ that minimizes the Integrated version of the Extended Cross-Validation quantity:

$$\mathrm{IECV}(h) = \int_{[c,d]} \mathrm{ECV}(y;h)dy, \tag{2.6}$$

and obtain the "*ECV bandwidth selector*"

$$\hat{h}_{p,\nu}^{\mathrm{ECV}} = \mathrm{adj}_{p,\nu}\hat{h}.$$

The integration in (2.6) has also a stabilizing effect on the variability of the ECV-quantity. The performance of this estimation procedure was investigated via four simulated examples in Section 5. We only present the results for Examples 2 and 4 (see Figures 2.c and 4.c) since Examples 1 and 3 show approximately the same performance as Example 2. As can be seen from those examples, the performance is good but an improvement is desirable. See Figure 4.c which shows a large variability of the bandwidth selection rule. Important gains can be obtained via a refinement, for which the basics are described in the next section. This refinement does not only improve the rate of convergence in the bandwidth selection procedure (compare Figure 2.b with 2.c, and Figure 4.b with 4.c), but also gives a better estimated curve in visual sense.

3 Assess of bias and variance in local polynomial fitting

The bias and variance in (1.3) are not directly accessible, since they depend on the unknown quantities, the residual $\mathbf{r}$ and the diagonal matrix Σ. Good finite sample estimates of the bias and variance are desirable in order to open a gate to a bandwidth selection procedure with a good overall performance.

The bias in (1.3) can clearly be approximated by $(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\tau$, where the i^{th} element of the $n \times 1$ vector τ equals

$$\beta_{p+1}(X_i - x_0)^{p+1} + \cdots + \beta_{p+a}(X_i - x_0)^{p+a}.$$

The choice of a is discussed below. With

$$s_{n,j} = \sum_{i=1}^{n}(X_i - x_0)^j K\left(\frac{X_i - x_0}{h}\right) \tag{3.1}$$

the approximated bias is equivalent with

$$S_n^{-1}\begin{pmatrix} \beta_{p+1}s_{n,p+1}+\cdots+\beta_{p+a}s_{n,p+a} \\ \vdots \\ \beta_{p+1}s_{n,2p+1}+\cdots+\beta_{p+a}s_{n,2p+a} \end{pmatrix}, \tag{3.2}$$

where $S_n = \mathbf{X}^T\mathbf{W}\mathbf{X}$ is the $(p+1)\times(p+1)$ matrix whose $(i,j)^{th}$ element is $s_{n,i+j-2}$. Thus, the bias vector can be estimated by

$$S_n^{-1}\begin{pmatrix} \hat{\beta}_{p+1}s_{n,p+1}+\cdots+\hat{\beta}_{p+a}s_{n,p+a} \\ \vdots \\ \hat{\beta}_{p+1}s_{n,2p+1}+\cdots+\hat{\beta}_{p+a}s_{n,2p+a} \end{pmatrix}, \tag{3.3}$$

where $\hat{\beta}_{p+1},\cdots,\hat{\beta}_{p+a}$ are the estimated regression coefficients from fitting a $(p+a)^{th}$ order polynomial locally.

Further, we set

$$s_{n,p+a+1}=0,\cdots,s_{n,2p+a}=0,$$

in order to reduce the effect of collinearity. See Fan and Gijbels (1993) for details.

The choice $a=4$ guarantees that the proposed selection procedure will be $\sqrt{n}$-consistent. On the other hand, the choice $a=2$ leads to a reduction of the computational efforts, while still having a selection rule which is not far from being $\sqrt{n}$-consistent. This makes this latter case attractable from practical point of view. Throughout the rest of the paper we will put $a=2$ for ease of presentation.

The variance in (1.3) can be approximated by

$$(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{W}^2\mathbf{X})(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\sigma^2(x_0), \tag{3.4}$$

using the local homoscedasity. Substitution of $\sigma^2(x_0)$ by a natural estimator — e.g. a residual sum of squares — leads to the variance estimator

$$(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{W}^2\mathbf{X})(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\hat{\sigma}^2(x_0) \tag{3.5}$$

where $\hat{\sigma}^2(x_0)$ is the weighted residual sum of squares from a $(p+a)^{th}$ order polynomial fit.

Now, the Mean Squared Error (MSE) of $\hat{\beta}_\nu(x_0) = \hat{m}^{(\nu)}(x_0)/\nu!$ is estimated by

$$\widehat{\mathrm{MSE}}_{p,\nu}(x_0; h) = \hat{b}^2_{p,\nu}(x_0) + \hat{V}_{p,\nu}(x_0),$$

where $\hat{b}_{p,\nu}(x_0)$ estimates the bias and is given by the $(\nu+1)^{th}$-element of the vector in (3.3). The variance estimator $\hat{V}_{p,\nu}(x_0)$ equals the $(\nu+1)^{th}$ diagonal element of the matrix in (3.5). This estimated MSE will serve as a second building block in the final bandwidth selection procedure.

Note that preliminary estimates $\hat{\beta}_{p+1}, \hat{\beta}_{p+2}, \cdots, \hat{\beta}_{p+a}$, and $\hat{\sigma}^2(x_0)$, based on a preliminary choice of h (to be specified in the next section via the ECV-criteria), are used to compute $\widehat{\mathrm{MSE}}_{p,\nu}(x_0; h)$.

4 Implementation

The material established in Sections 2 and 3 enables us to develop an appealing methodology for selecting a constant or variable bandwidth. Suppose the interest is in estimating $m^{(\nu)}(x_0)$ by using a p^{th} order polynomial. Usually $p = \nu + 1$ (see Fan and Gijbels (1993)). The selection rules for each type of bandwidth are presented below.

4.1 Bandwidth selection rules

Constant bandwidth

The proposed bandwidth selection rule reads as follows. Fit a polynomial of order $p + 2$, use IECV in (2.6) to select the optimal bandwidth for estimating β_{p+1} and obtain the estimates $\hat{\beta}_{p+1}(x_0), \hat{\beta}_{p+2}(x_0)$ and $\hat{\sigma}^2(x_0)$. Now find the bandwidth that minimizes the estimated Integrated Mean Squared Error:

$$\hat{h}^{\mathrm{R}}_{p,\nu} = \arg\min_h \int_{[c,d]} \widehat{\mathrm{MSE}}_{p,\nu}(y; h)dy,$$

and use this to fit a polynomial of order p . Throughout the paper we will refer to this particular "plug-in" bandwidth selector as "*Refined bandwidth selector*".

The above refinement of the previously described selection rule does lead to a considerable improvement, as evidenced by the examples in Section 5. See Figures 2 and 4, a — c. If, in a particular situation, the *one-stage* procedure — the ECV bandwidth selector — appears to be of sufficient performance, then one can stick to this since it is computationally less involved. However, computation times for the "ECV-procedure" as well as for the "Refined" (*two-stage*) procedure are very fast. Therefore, if computation is no issue, we recommend to use the Refined bandwidth selector since its performance is superior.

Remark that the proposed estimation procedure does not require the choice of any parameter, and hence is fully automatic.

Variable bandwidth

The experience with the constant bandwidth choice showed that a refined procedure is recommendable. Selecting a variable bandwidth is even more involved and hence a similar kind of procedure is a minimum requirement. The above exposed ideas are now used to establish a selection procedure for a variable bandwidth. The main difference is here that we start with splitting up the interval of estimation $[c, d]$ in subintervals, say I_k, and use the Refined bandwidth selector in each interval. In detail this procedure reads as:

> For each interval I_k, fit a polynomial of order $p+2$ and select the optimal bandwidth for estimating β_{p+1} by minimizing $\text{IECV}(h) = \int_{I_k} \text{ECV}(y; h)dy$. Smooth the resulting bandwidth stepfunction by averaging locally, using the same smoothing parameter as for the initial partition, i.e. the length of I_k. Use this smoothed bandwidth function to fit a polynomial of order $p + 2$, and obtain $\hat{\beta}_{p+1}(x_0), \hat{\beta}_{p+2}(x_0)$, and $\hat{\sigma}^2(x_0)$.
> For each interval I_k, choose the bandwidth which minimizes the estimated Integrated Mean Squared Error $\int_{I_k} \widehat{\text{MSE}}(y; h)dy$. Smooth the resulting bandwidth stepfunction, using again the length of I_k, and fit a polynomial of order p.

We remark that the smoothing step in the above procedure leads to a smoother estimated curve. In our simulated examples, we split the interval $[c, d]$ into $[n/10 \log(n)]$ pieces. Such a choice reflects somewhat the availability of data for exploiting complex structures.

4.2 Practical implementation

We would like to make some remarks on the practical implementation of the estimation procedure. First of all, in practice the estimated curves are evaluated in grid points x_j, $j = 1, \cdots, n_{\mathrm{grid}}$. Consequently, the integrals involved in the methodology are implemented as averages over appropriate grid points.

The methodology involves a few minimization problems. The functions in h which have to be minimized are of a very complicated form, and hence usage of the Newton-Raphson method for finding a minimum is almost impossible. A feasible method is to compare function values at grid points (typically of geometric type). Suppose we want to minimize a function $M(h)$ over an interval $[h_{\mathrm{min}},\ h_{\mathrm{max}}]$. Here $M(h)$ could either be IECV or $\int \widehat{\mathrm{MSE}}(y; h)dy$. Starting from $h = h_{\mathrm{min}}$, keep inflating h by a factor C and compute $M(h)$ at these geometric grid points. Stop when the function values $M(h)$ increase *consecutively* a certain number of times, say IUP, or when $h > h_{\mathrm{max}}$. Now, choose the minimizer of $M(h)$ as the grid point having the smallest computed $M(h)$ value.

Fitting a local polynomial at a large bandwidth is computationally very costly. With the above minimization procedure we try to avoid a fit with a large bandwidth, unless it is absolutely necessary. In our implementation we took $h_{\mathrm{min}} = (X_{(n)} - X_{(1)})/n,\ h_{\mathrm{max}} = (X_{(n)} - X_{(1)})/2, \mathrm{IUP} = 3, C = 1.1$. With those choices, the described minimization method enables us to compute an estimated curve with fully automatically selected bandwidth for sample size $n = 200$ in less than 10 seconds using a Sparc 2 workstation.

Finally, it should be mentioned that there are possibilities for improving the computational speed. Fast computation algorithms such as linear binning and updating could be implemented. A thorough investigation of fast implementations of nonparametric curve estimators was carried through by Fan and Marron (1993) and Wand (1993).

5 Test examples

We now investigate the performance of the proposed methodology, via a variety of simulated examples, and the Motorcycle Data (see e.g. Härdle (1990)). The study concerns estimation of $m(\cdot)$ as well as its derivatives. In each of the examples we use the Epanechnikov kernel. The number of simulations is 400. The table below summarizes the models used in the simulated examples, and indicates the choices of $m(x)$ and $\sigma(x) \equiv \sigma$ in the general regression model $Y = m(X) + \sigma(X)\varepsilon$. The table also lists an approximation of the noise to signal ratio $\sigma^2/(\mathrm{Var}(m(X)) + \sigma^2)$ (see column 4), which is an indicator for the difficulty of the estimation problem. The bigger this ratio the harder the problem. Note that for the first four simulated examples the noise to signal ratio is very high, which implies that for those examples the estimation task is difficult. For Examples 5 — 8, the noise to signal ratio is more moderate which lightens the estimation task.

Table 2: simulated examples

Example	$m(x)$	σ	$\approx$ signal/noise	sample size
1	$x + 2e^{-16x^2}$	0.4	1/3	50 200 800
2	$\sin(2x) + 2e^{-16x^2}$	0.3	1/3	50 200 800
3	$0.3e^{-4(x+1)^2} + 0.7e^{-16(x-1)^2}$	0.1	1/2	50 200 800
4	$0.4x + 1$	0.15	1/3	50 200 800
5	$24\sqrt{x(1-x)}\sin(2\pi 1.05/(x+0.05))$	1.0	1/7	2048
6	(see D&J (1992))	1.0	1/7	2048
7	(see D&J (1992))	1.0	1/7	2048
8	(see D&J (1992))	1.0	1/7	2048

* Donoho and Johnstone (1992) was abbreviated as D&J (1992).

In Example 9 we analyse the Motorcycle Data.

For Examples 1 — 4 we used a random uniform design, i.e. $X \sim \mathrm{Uniform}(-2, 2)$. For Examples 5 — 9 the fixed uniform design $x_i = \frac{i}{n}$ was applied. The estimated curve is calculated in grid points. The number of grid points is 101 for Examples 1 — 4 and 9, and for Examples 5 — 8 we took $n_{\mathrm{grid}} = 1001$. In each of the examples we do local linear fits ($p = 1$), and take $a = 2$ (see (3.2)).

5.1 Constant bandwidth

The performance of the constant bandwidth selection procedure is illustrated via Examples 1 — 4. For each of the examples we provide two pictures. A first picture presents the true regression curve, a typical simulated data set ($n = 200$) and some representative estimated curves based on 400 simulations. Those representatives were chosen as follows: for each estimated curve compute the Mean Squared Error averaged over all grid points, rank all estimated curves according to this measure, and select the estimated curves corresponding to the $10th\%$, the $50th\%$ and the $90th\%$ rank-observation. This first picture gives a visual impression of the quality of the estimated regression curve $\hat{m}(x)$.

A second picture reports on the relative orders of the estimated bandwidth

$$(\hat{h}_{\nu,opt} - h_{\nu,opt})/h_{\nu,opt} \tag{5.1}$$

where $h_{\nu,opt}$ is the theoretical optimal constant bandwidth, computed via (1.3). The 400 relative errors are summarized by means of a kernel density estimate. For this kernel density estimator we used a Gaussian kernel and Silverman's (1986) bandwidth selector $h = 1.06\ s\ n^{-0.2}$, where s denotes the standard deviation of the data. The three bars in the second picture represent the percentage of relative errors $(\hat{h}_{\nu,opt} - h_{\nu,opt})/h_{\nu,opt}$ less than 20 % for sample sizes 50, 200 and 800. In Example 4 we took $h_{\nu,opt} = 1$, and reported the percentage of selected bandwidths using more than 40 % of the data, i.e: $\hat{h}_{\nu,opt} > 0.8$.

The more spiky the curve of relative errors, the better the criteria for selecting the bandwidth. From the pictures it is clear that the estimated bandwidth converges to its theoretical counterpart (the curves are more spiky when sample size increases).

Figures 1 — 4, a and b report on the performance of the Refined bandwidth selector.

Put Figures 1 — 4, a and b about here

Figures 1.a — 4.a: *A typical simulated data set along with 3 representative estimated curves (n =* 200*). Solid line: true regression function; dashed lines: 3 representative estimated curves.*

Figures 1.b — 4.b: *Kernel density estimates for the relative errors of the Refined bandwidth selector. The 3 curves represent the kernel density estimates for the distribution of* *the relative errors (5.1), normalized to have maximum height 1. The 3 vertical bars show the percentage of relative errors less than 20 % (from left to right for* $n = 50, 200, 800$*).*

For comparison purposes we present the results of the one-stage procedure with the ECV bandwidth selector, for Examples 2 and 4 in respectively Figures 2.c and 4.c. The performance of the ECV procedure was very good for Examples 1 — 3, while for Example 4 the selected bandwidth ended up with having a large variability. Moreover, a further conclusion can be drawn from the comparison of Figures 2.b and 2.c (and similar pictures for Examples 1 and 3 not presented here): one can see that the Refined bandwidth selector has a faster convergence rate (is getting 'spiky' faster) than the ECV bandwidth selector.

Put Figures 2.c and 4.c about here

Figures 2.c and 4.c: *Kernel density estimates for the relative errors of the ECV bandwidth selector. The 3 curves represent the kernel density estimates for the distribution of the relative errors (5.1), normalized to have maximum height 1. The 3 vertical bars show the percentage of relative errors less than 20 % (from left to right for* $n = 50, 200, 800$*).*

5.2 Local variable bandwidth

We first study the performance of the procedure with variable bandwidth when estimating the regression function $m(\cdot)$ itself. We do this for Examples 1 — 9. For each of the Examples 1 — 8 we present a picture that summarizes the 400 estimated curves via the percentiles, as before.

The sample size for Examples 1 — 4 was 200. The results for those examples are given in Figures 1.c, 2.d, 3.c and 4.d, which all indicate that the data-driven choice of the variable bandwidth does a good job.

Put Figures 1.c, 2.d, 3.c and 4.d about here

Figures 1.c, 2.d, 3.c and 4.d: *Three representatives of the estimated curves for sample size* $n = 200$*, based on 400 simulations.*

Of particular interest are the analysis of Examples 5 — 8, reflected in Figures 5 — 8, a and b. The proposed methodology captures very nicely the complexity of each of the curves, due clearly to the appropriate data-driven choice of the variable bandwidth. In comparing the performance of our methododology with that of the Wavelets-packages provided in Donoho and Johnstone (1992), the present method performs at least as good. In other words, the spatial-adaptation property of Wavelets can also easily be achieved via local polynomial fitting, using an appropriate variable bandwidth. Moreover, a variable bandwidth possesses the flexibility of adapting the smoothing parameter to the location point.

Put Figures 5 — 8, a and b about here

Figures 5.a – 8.a : *A typical simulated data set with sample size* $n = 2048$.

Figures 5.b – 8.b : *A typical (with median MISE) estimated curve, based on 31 simulations.*

Finally, we present the analysis of the Motorcycle data in Figure 9.

Put Figure 9 about here

Figure 9: *Motorcycle data and its estimated curve.*

We next report on the performance of the procedure for estimating the derivative curve $m'(\cdot)$. For this illustration we consider Examples 2 and 5. The results for Example 2 are in Figure 2.e, which gives the true derivative curve and typical estimated derivative curves for sample sizes 200 and 800. Figure 5.c presents the true derivative curve and a typical estimated derivative curve for Example 5. Both examples demonstrate that the procedure also works out very neatly for derivative estimation.

Put Figures 2.e and 5.c about here

Figure 2.e: *Derivative estimation. Solid line: the true derivative function; Dotted line: the estimated derivative curve for* $n = 200$*; Dashed line: the estimated derivative curve for* $n = 800$.

Figure 5.c: *Derivative estimation. Solid line: the true derivative function; Dotted line: the estimated derivative curve (*$n = 2048$*).*

6 Justification of the proposed method

In this section we give the derivations which go with the results presented in Section 2. Let us first of all look at appropriate approximations for the conditional bias and variance in (1.3).

6.1 Assess to bias and variance

Starting from definition (3.1) it is easy to see that

$$s_{n,j} = f_X(x_0) s_j n h^{j+1}(1 + O_P(h)), \tag{6.1}$$

and hence as a consequence

$$S_n = \mathbf{X}^T \mathbf{W} \mathbf{X} = f_X(x_0) nh HSH(1 + O_P(h)), \tag{6.2}$$

where $H = \operatorname{diag}(1, h, \cdots, h^p)$. Therefore a further approximation of the bias in (3.2) is provided by

$$h^{p+1} H^{-1} S^{-1} H^{-1} \begin{pmatrix} \beta_{p+1} s_{p+1} + O_P(h) \\ \beta_{p+1} h s_{p+2} + O_P(h^2) \\ \vdots \\ \beta_{p+1} h^p s_{2p+1} + O_P(h^{p+1}) \end{pmatrix}.$$

This also leads to

$$\mathrm{E}(\hat{\beta}_\nu | X_1, \cdots, X_n) = h^{p+1-\nu} \beta_{p+1} e_{\nu+1}^T S^{-1} \begin{pmatrix} s_{p+1} \\ \vdots \\ s_{2p+1} \end{pmatrix} + o_P(h^{p+1-\nu}),$$

where $e_{\nu+1}$ denotes the unit $(p+1) \times 1$ vector, containing 1 on the $(\nu+1)^{th}$ position. This provides the bias part in (2.1).

For the conditional variance we proceed as follows. Using similar arguments as before, we find that

$$S_n^* = \mathbf{X}^T \mathbf{W}^2 \mathbf{X} = f_X(x_0) nh HS^*H(1 + O_P(h)).$$

This together with (6.2) leads to a further approximation of the conditional variance in (3.4), namely,

$$\frac{1}{nhf_X(x_0)} H^{-1}S^{-1}S^*S^{-1}H^{-1}\sigma^2(x_0),$$

and for example,

$$\mathrm{Var}(\hat{\beta}_\nu | X_1, \cdots, X_n) = a_\nu \frac{\sigma^2(x_0)}{f_X(x_0)} \frac{1}{nh^{1+2\nu}} + o_P(\frac{1}{nh^{1+2\nu}}), \qquad (6.3)$$

which provides the variance part in (2.1).

6.2 Proof of Theorem 1

Note first of all that for the weighted residual sum of squares, defined in (2.3), we have

$$\begin{aligned} \hat{\sigma}^2(x_0) &= \frac{1}{\mathrm{tr}(\mathbf{W} - \mathbf{WX}(\mathbf{X}^T\mathbf{WX})^{-1}\mathbf{X}^T\mathbf{W})} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 W_i \\ &= \frac{1}{\mathrm{tr}(\mathbf{W} - \mathbf{WX}(\mathbf{X}^T\mathbf{WX})^{-1}\mathbf{X}^T\mathbf{W})} (\mathbf{y} - \mathbf{X}\hat{\beta})^T \mathbf{W} (\mathbf{y} - \mathbf{X}\hat{\beta}) \\ &= \frac{1}{\mathrm{tr}(\mathbf{W} - \mathbf{WX}(\mathbf{X}^T\mathbf{WX})^{-1}\mathbf{X}^T\mathbf{W})} \mathbf{y}^T(\mathbf{W} - \mathbf{WX}(\mathbf{X}^T\mathbf{WX})^{-1}\mathbf{X}^T\mathbf{W})\mathbf{y}. \end{aligned}$$

and consequently,

$$\begin{aligned} & \mathrm{E}(\hat{\sigma}^2(x_0) | X_1, \cdots, X_n) \\ &= \frac{1}{\mathrm{tr}(\mathbf{W} - \mathbf{WX}(\mathbf{X}^T\mathbf{WX})^{-1}\mathbf{X}^T\mathbf{W})} \mathbf{m}^T(\mathbf{W} - \mathbf{WX}(\mathbf{X}^T\mathbf{WX})^{-1}\mathbf{X}^T\mathbf{W})\mathbf{m} \\ & \quad + \sigma^2(x_0), \qquad (6.4) \end{aligned}$$

where we used the local homoscedasity. Approximating $\mathbf{r} = \mathbf{m} - \mathbf{X}\beta$ by

$$r_i = m(X_i) - \sum_{j=0}^p \beta_j (X_i - x_0)^j = \beta_{p+1}(X_i - x_0)^{p+1} + O_P(h^{p+2}),$$

the first term on the right-hand side of (6.4) becomes

$$\begin{aligned} & \frac{1}{\mathrm{tr}(\mathbf{W} - \mathbf{WX}(\mathbf{X}^T\mathbf{WX})^{-1}\mathbf{X}^T\mathbf{W})} \mathbf{r}^T(\mathbf{W} - \mathbf{WX}(\mathbf{X}^T\mathbf{WX})^{-1}\mathbf{X}^T\mathbf{W})\mathbf{r} \\ &= \frac{(s_{n,2p+2} - (s_{n,p+1}, \cdots, s_{n,2p+1}) S_n^{-1} (s_{n,p+1}, \cdots, s_{n,2p+1})^T)}{\mathrm{tr}(\mathbf{W} - \mathbf{WX}(\mathbf{X}^T\mathbf{WX})^{-1}\mathbf{X}^T\mathbf{W})} \beta_{p+1}^2 (1 + o_P(h)). \qquad (6.5) \end{aligned}$$

Further, remark that

$$\begin{aligned}\mathrm{tr}(\mathbf{W}-\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}) &= s_{n,0}-\mathrm{tr}((\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^2\mathbf{X}) \\ &= f_X(x)s_0nh+O_P(1+nh^3). \qquad (6.6)\end{aligned}$$

Finally, as provided in (6.3),

$$V_{n,0}=a_0\frac{1}{f_X(x_0)}\frac{1}{nh}+o_P(\frac{1}{nh}). \qquad (6.7)$$

Using (6.1) — (6.7) we find that

$$\begin{aligned}&\mathrm{E}(\hat{\sigma}^2(x_0)|X_1,\cdots,X_n) \\ &= \frac{(s_{n,2p+2}-(s_{n,p+1},\cdots,s_{n,2p+1})S_n^{-1}(s_{n,p+1},\cdots,s_{n,2p+1})^T)}{\mathrm{tr}(\mathbf{W}-\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W})}\beta_{p+1}^2+\sigma^2(x_0)+o_P(h^{2p+2}) \\ &= C_p\beta_{p+1}^2h^{2p+2}+\sigma^2(x_0)+o_P(h^{2p+2}),\end{aligned}$$

with C_p as in (2.5).

This together with (6.7) leads to

$$\begin{aligned}&\mathrm{E}(\mathrm{ECV}(x_0;h)|X_1,\cdots,X_n) \\ &= \mathrm{E}(\hat{\sigma}^2(x_0)|X_1,\cdots,X_n)\,(1+(p+1)V_{n,0}) \\ &= \sigma^2(x_0)+C_p\beta_{p+1}^2h^{2p+2}+(p+1)a_0\frac{\sigma^2(x_0)}{nhf_X(x_0)}+o_P(h^{2p+2}+\frac{1}{nh})\end{aligned}$$

which completes the proof. □

REFERENCES

Brockmann, M., Gasser, T. and Hermann, E. (1993). Locally adaptive bandwidth choice for kernel regression estimators. *Jour. Amer. Statist. Assoc.*, to appear.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Jour. Amer. Statist. Assoc.*, **74** 829 – 836.

Cleveland, W.S. and Devlin, S.J. (1988). Locally-weighted regression: an approach to regression analysis by local fitting. *Jour. Amer. Statist. Assoc.* **83** 597 – 610.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, **31**, 377 – 403.

Donoho, D.L. and Johnstone, I.M. (1992). Ideal spatial adaptation via wavelet shrinkage. *Techn. Report* , Department of Statistics, Stanford University.

Fan, J. (1992). Design-adaptive nonparametric regression. *Jour. Amer. Statist. Assoc.*, **87**, 998 - 1004.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiency. *Ann. Statist.*, **21**, 196 - 216.

Fan, J. , Gasser, T., Gijbels, I. , Brockman, M. and Engel, J. (1993). Local polynomial fitting: a standard for nonparametric regression. *Manuscript.*

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.*, **20**, 2008 - 2036.

Fan, J. and Gijbels, I. (1993). Bandwidth and adaptive order selection for local polynomial fitting in function estimation. *Manuscript.*

Fan, J. and Marron, J.S. (1993). Fast implementations of nonparametric curve estimators. Department of Statistics, University of North Carolina. *Mimeo Series # 2093.*

Gasser, T., Kneip, A. and Köhler, W. (1991). A flexible and fast method for automatic smoothing. *Jour. Amer. Statist. Assoc.*, **86**, 643 - 652.

Hall, P., Sheather, S.J., Jones, M.C. and Marron, J.S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika*, **78**, 263 - 271.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press, Boston.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models (with discussion). *Statist. Sci.*, **1**, 297 - 318.

Jones, M.C., Marron, J.S. and Sheather, S.J. (1992). Progress in data based bandwidth selection for kernel density estimation. Department of Statistics, University of North Carolina. *Mimeo Series # 2088.*

Müller, H.-G. (1988). *Nonparametric Analysis of Longitudinal Data.* Springer Verlag, Berlin.

Müller, H.-G. and Stadtmüller, U. (1987). Variable bandwidth kernel estimators of regression curves. *Ann. Statist.*, **15**, 182 - 201.

Ruppert, D. and Wand, M.P. (1992). Multivariate weighted least squares regression. *Tech. Report no. 92-4.* Department of Statistics, Rice University.

Ruppert, D., Sheather, S.J. and Wand, M.P. (1993). An effective bandwidth selector for local least squares regression. *Manuscript.*

Sheather, S.J. and Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc., Series B*, **53**, 683 - 690.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Stone, C.J. (1977). Consistent Nonparametric Regression. *Ann. Statist.*, **5**, 595 - 645.

Vieu, P. (1991). Nonparametric regression: optimal local bandwidth choice. *J. Roy. Statist. Soc., Series B*, **53**, 453 - 464.

Wahba, G. (1977). A survey of some smoothing problems and the method of generalized cross-validation for solving them. In *Applications of Statistics* (P.R. Krisnaiah, ed.), 507 - 523. Amsterdam: North Holland.

Wand, M.P. (1993). Fast computation of multivariate kernel estimators. *Manuscript.*

Ex 1: Typical data with estimated curves (n=200)

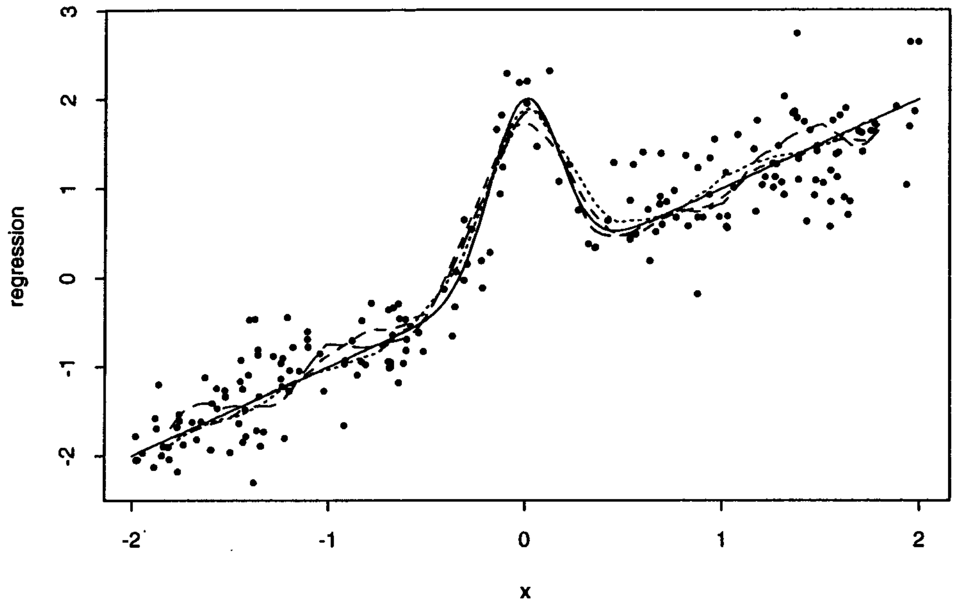

Figure 1.a

Ex 1 : Relative Errors of Refined Bandwidth Selector

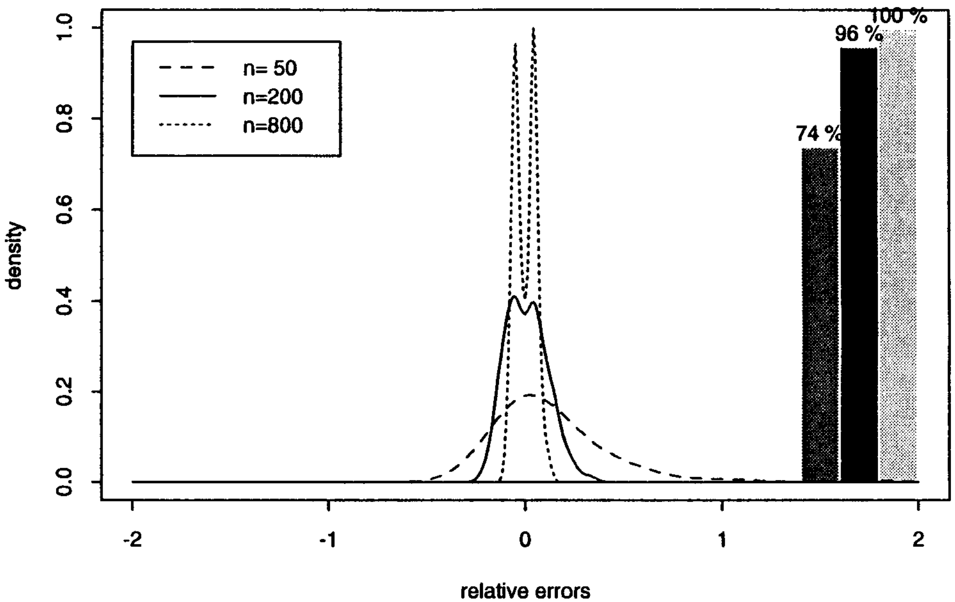

Figure 1.b

Ex 2: Typical data with estimated curves (n=200)

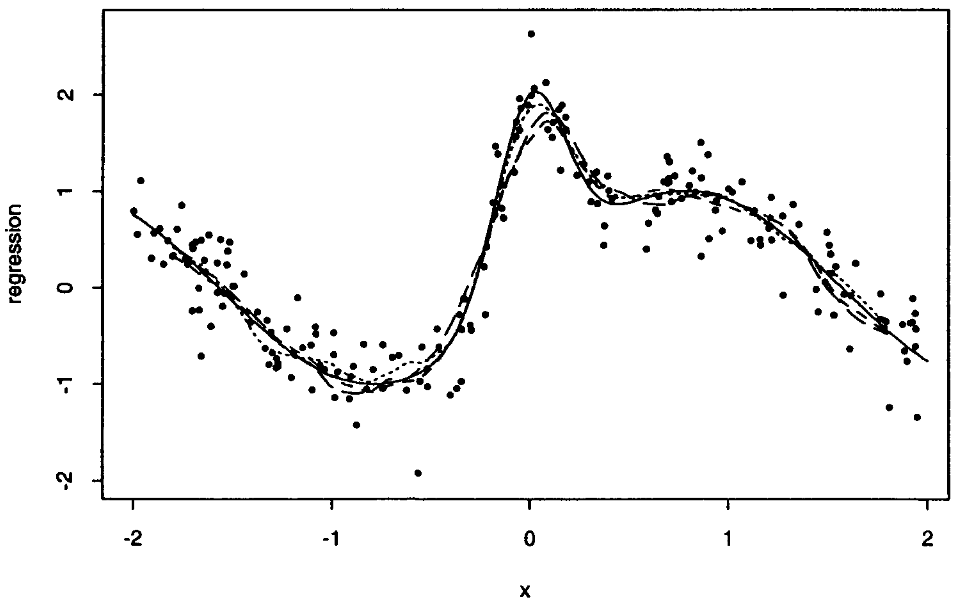

Figure 2.a

Ex 2 : Relative Errors of Refined Bandwidth Selector

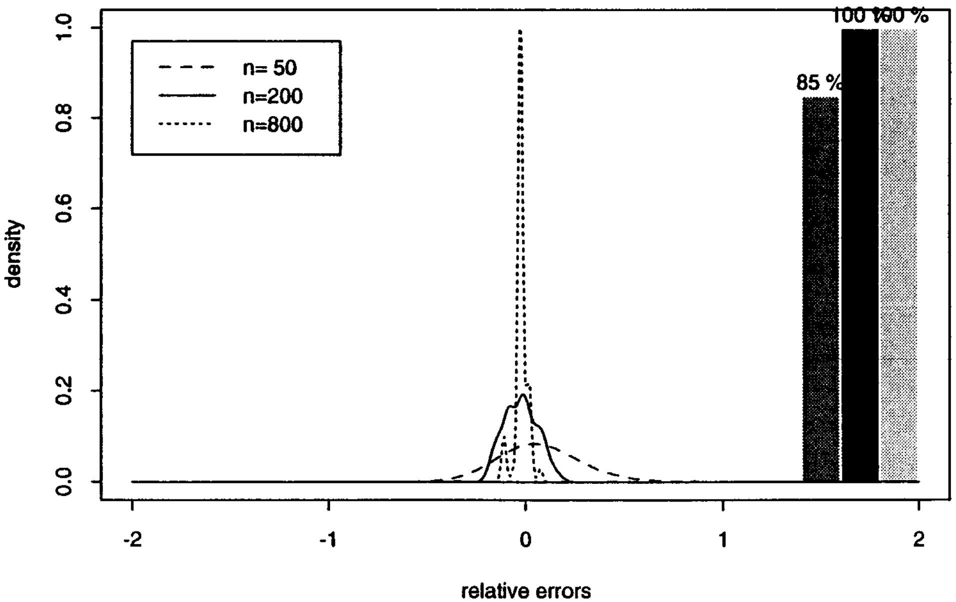

Figure 2.b

Ex 3: Typical data with estimated curves (n=200)

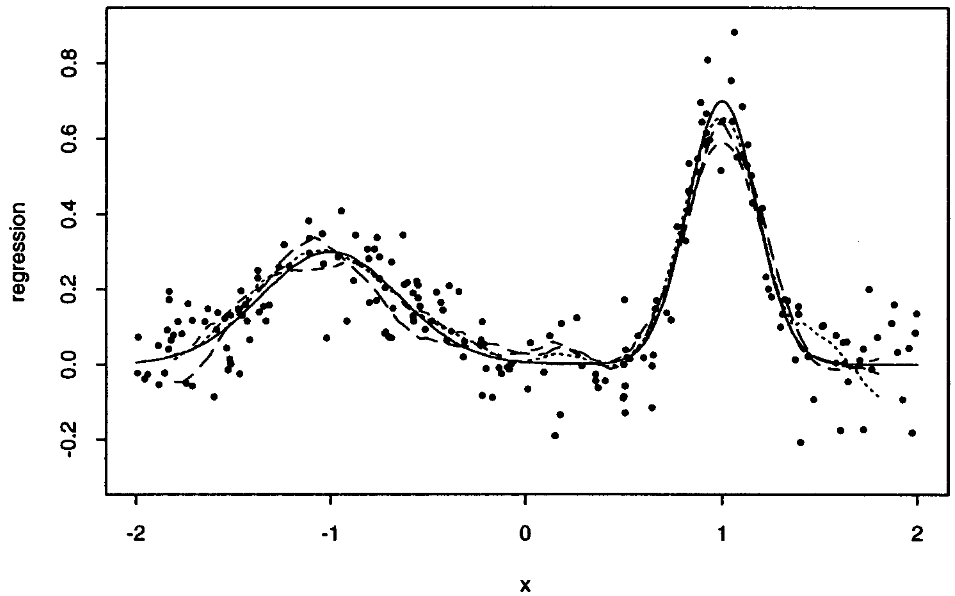

Figure 3.a

Ex 3 : Relative Errors of Refined Bandwidth Selector

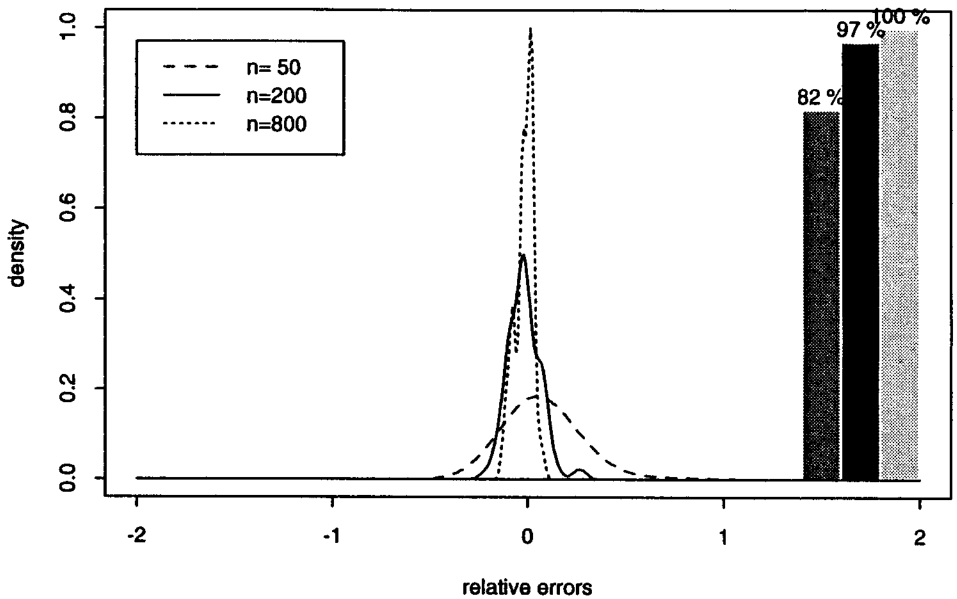

Figure 3.b

Ex 4: Typical data with estimated curves (n=200)

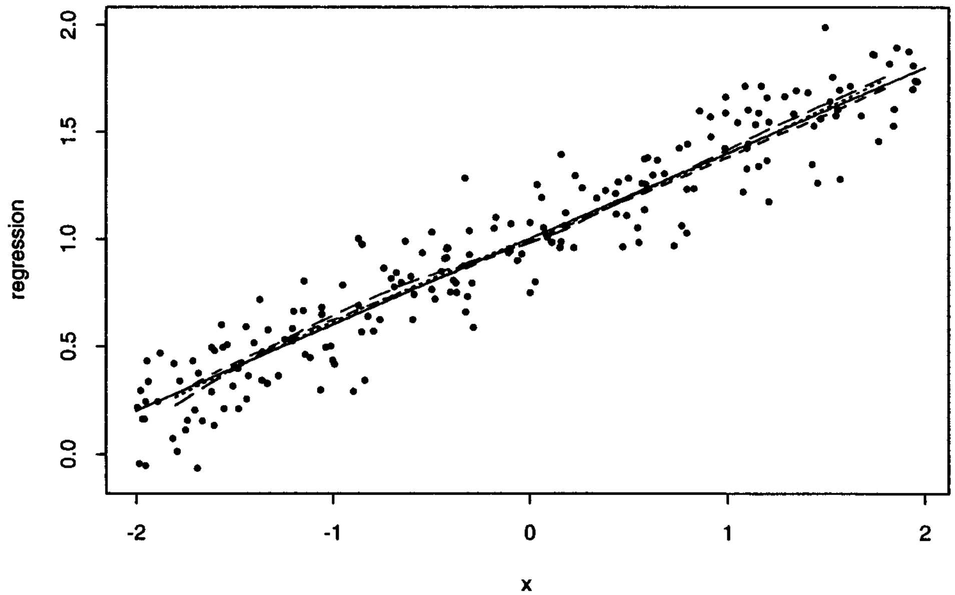

Figure 4.a

Ex 4 : Relative Errors of Refined Bandwidth Selector

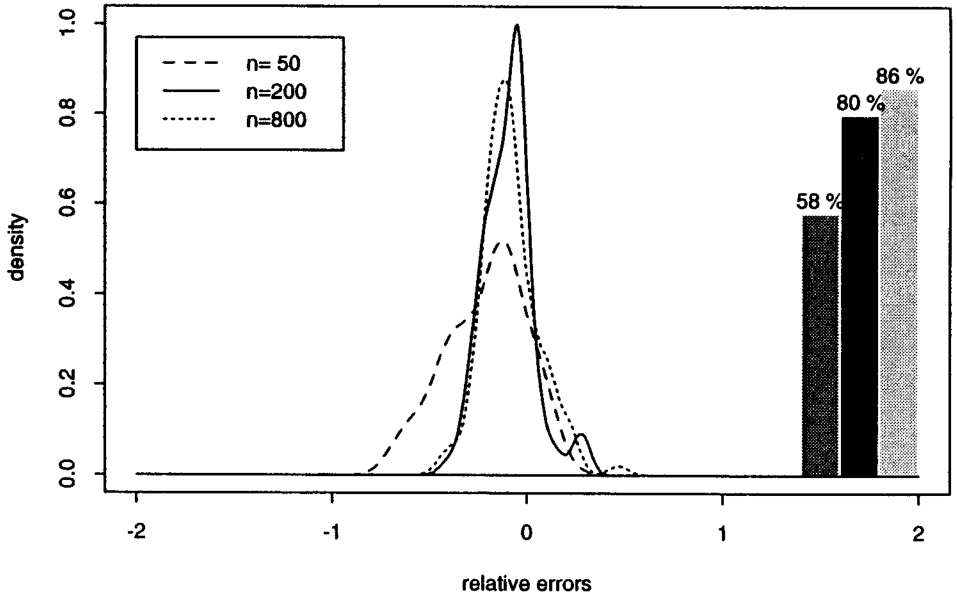

Figure 4.b

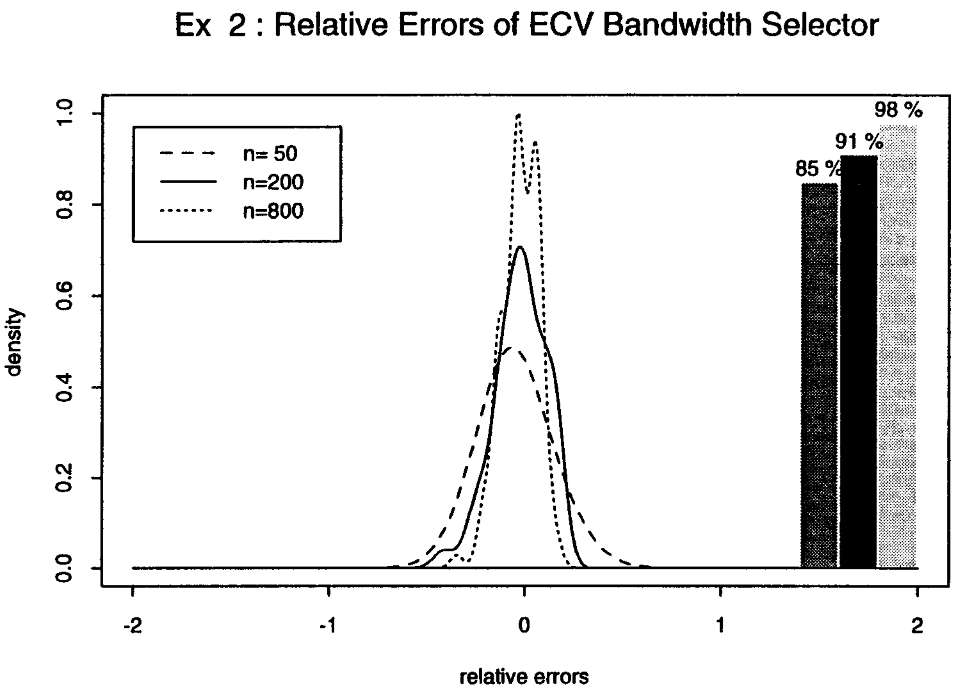

Figure 2.c

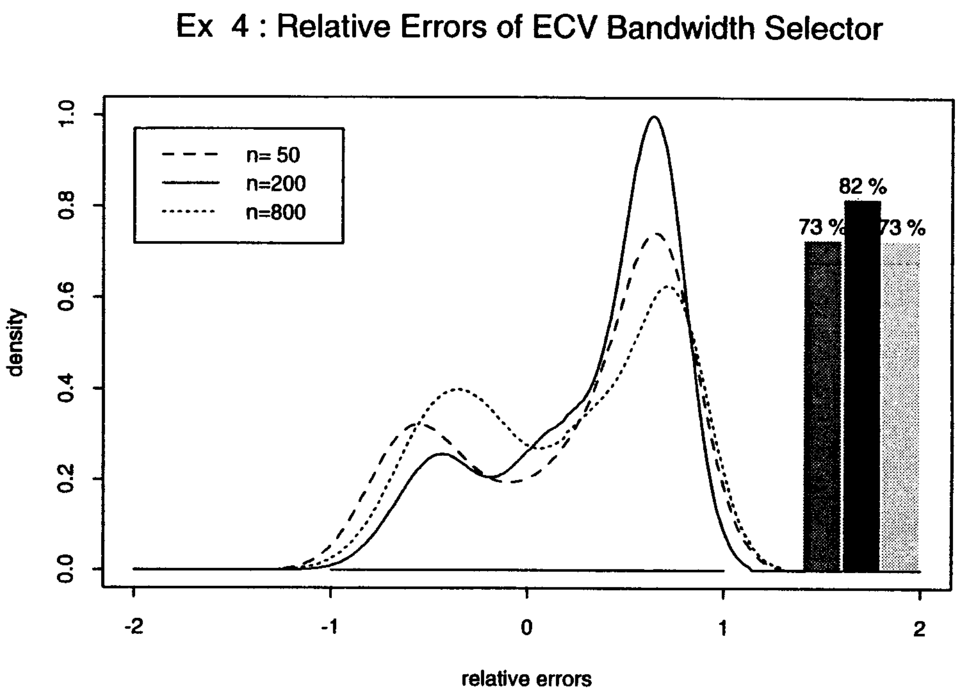

Figure 4.c

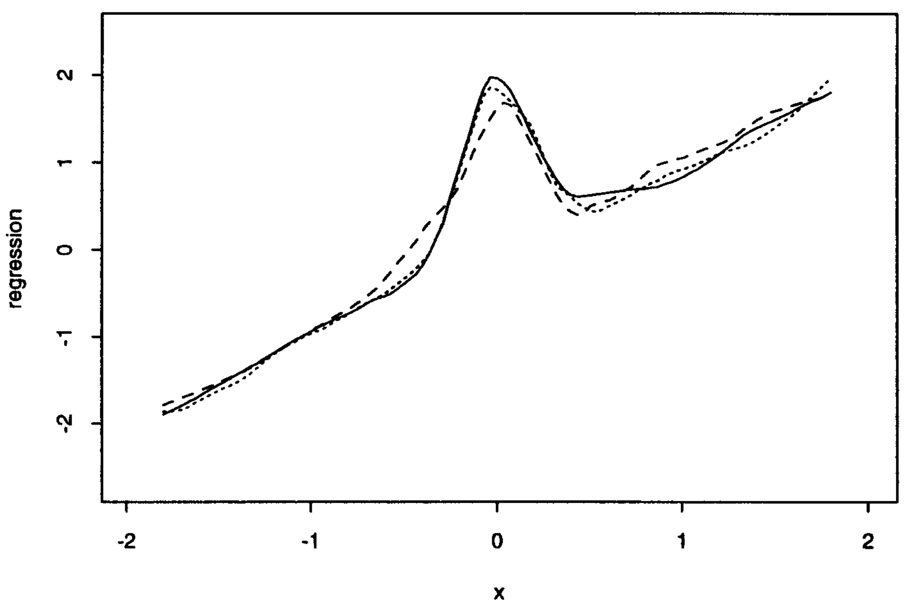
Ex 1: Typical estimated curves with n = 200

Figure 1.c

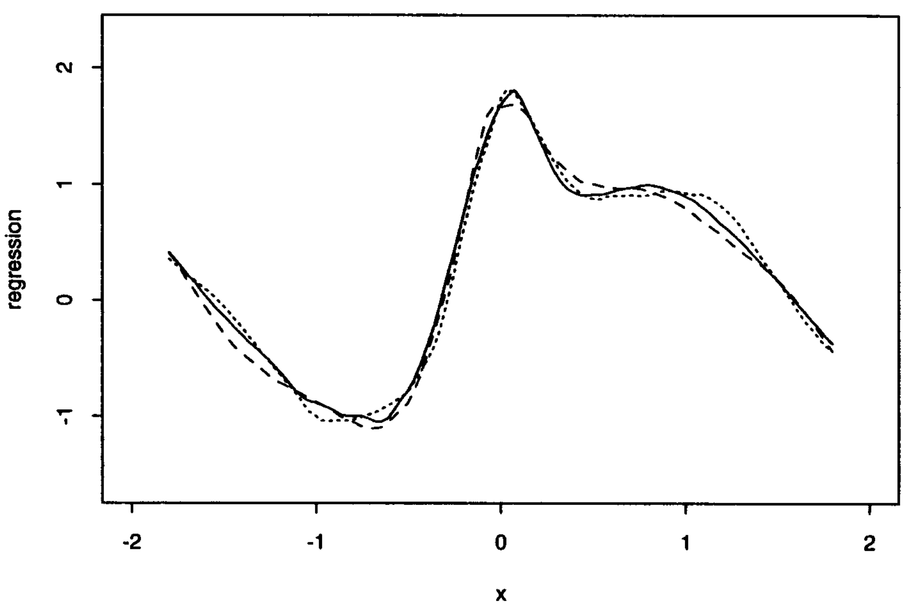
Ex 2: Typical estimated curves with n = 200

Figure 2.d

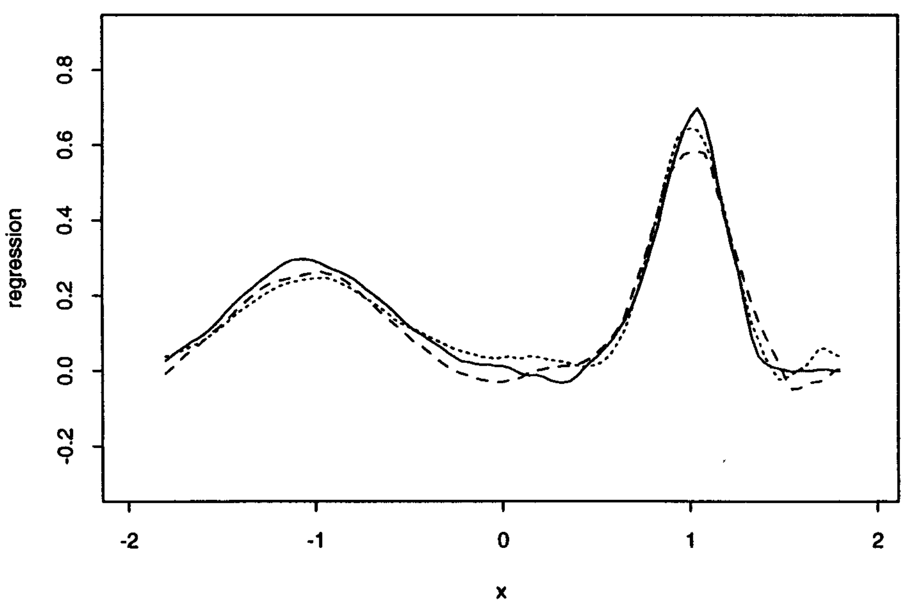
Ex 3: Typical estimated curves with n = 200

Figure 3.c

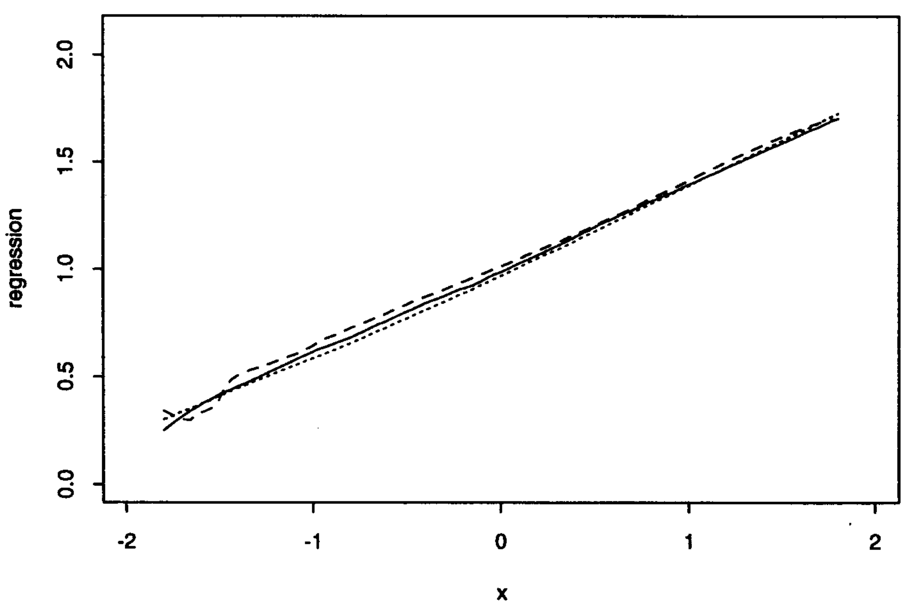
Ex 4: Typical estimated curves with n = 200

Figure 4.d

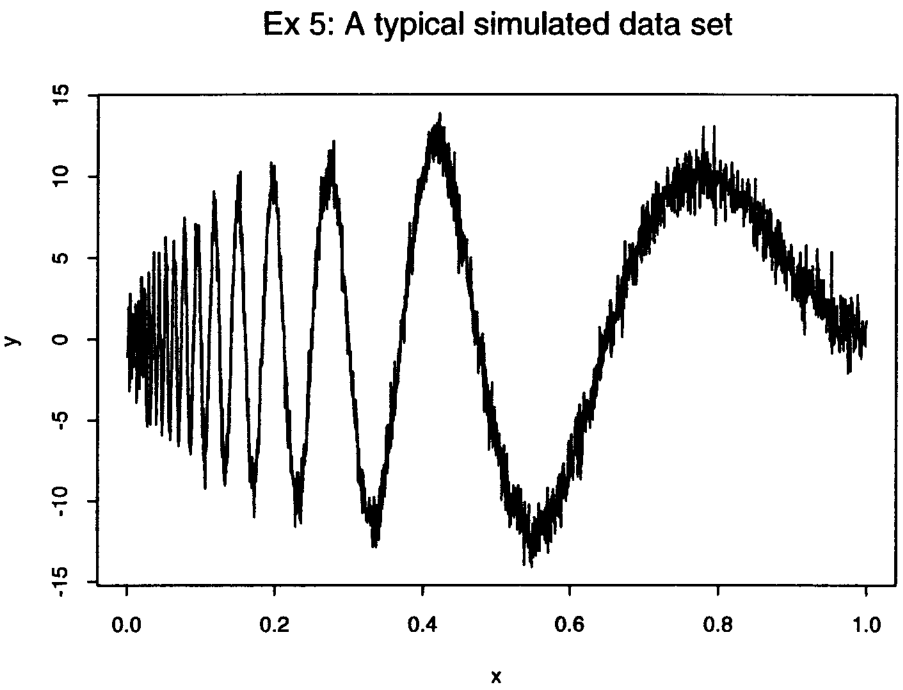
Ex 5: A typical simulated data set

Figure 5.a

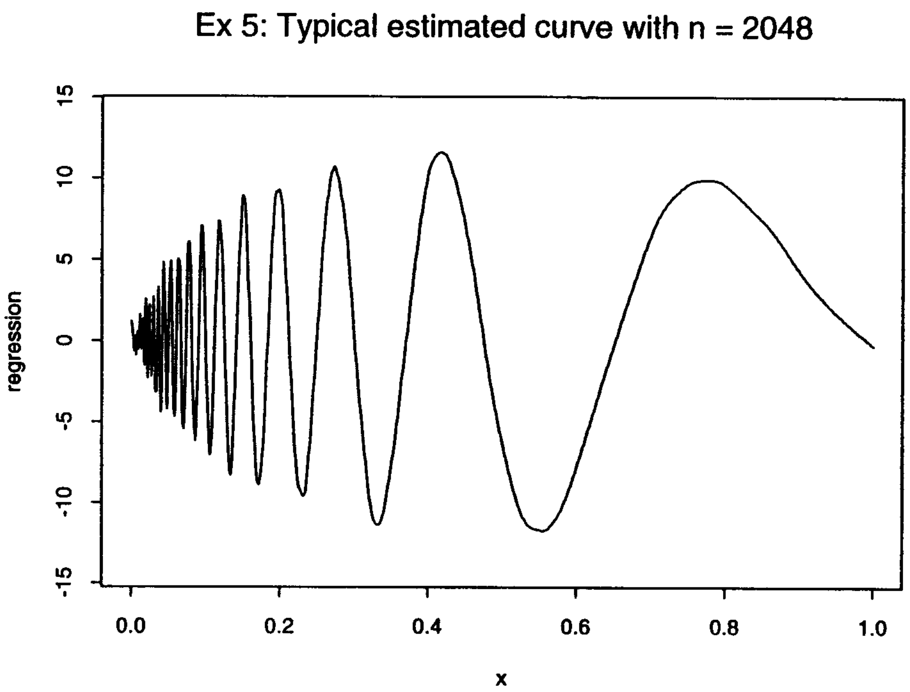
Ex 5: Typical estimated curve with n = 2048

Figure 5.b

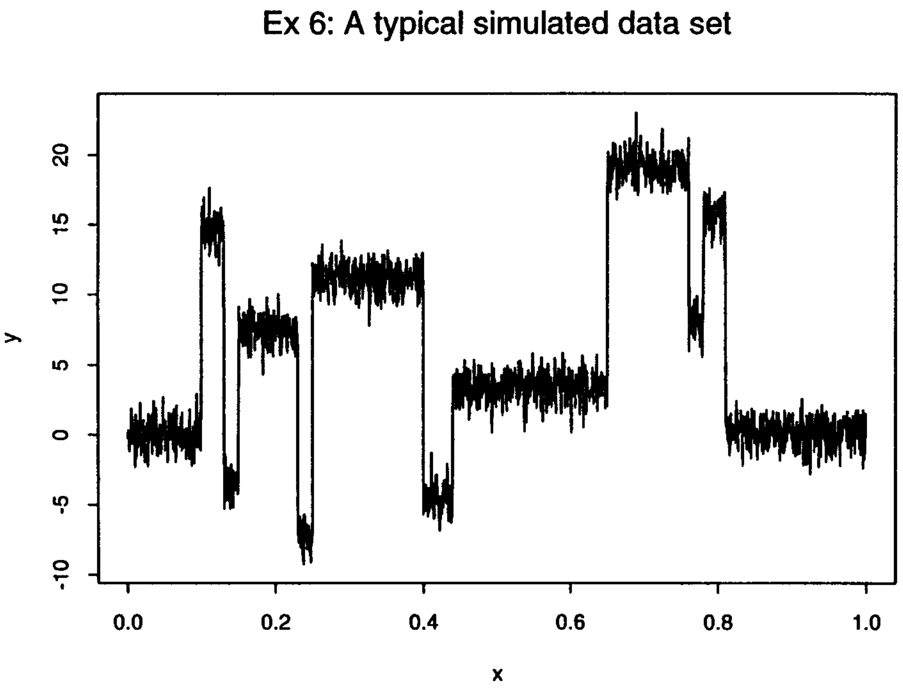
Ex 6: A typical simulated data set

Figure 6.a

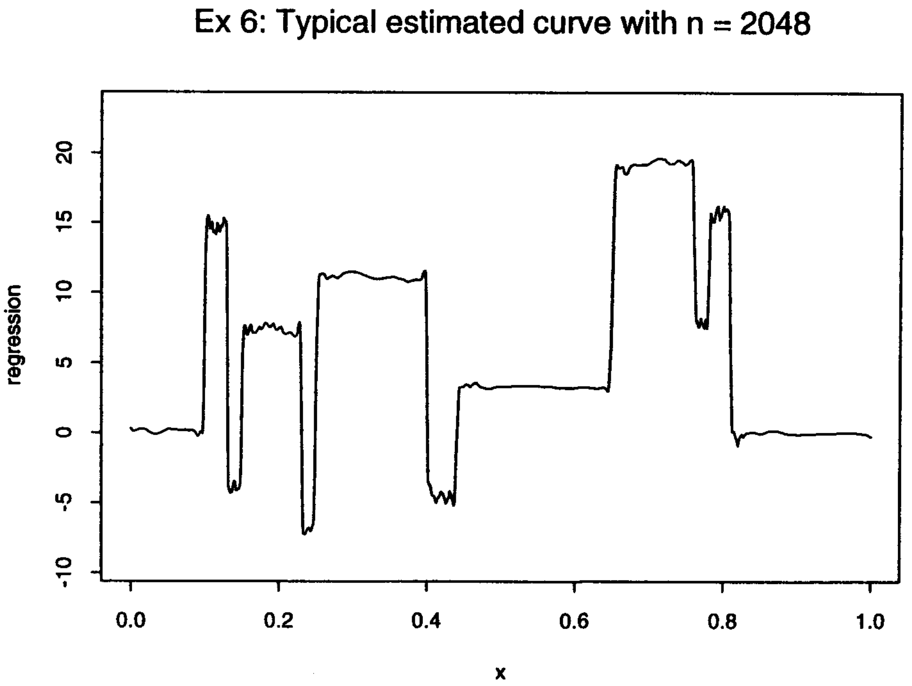
Ex 6: Typical estimated curve with n = 2048

Figure 6.b

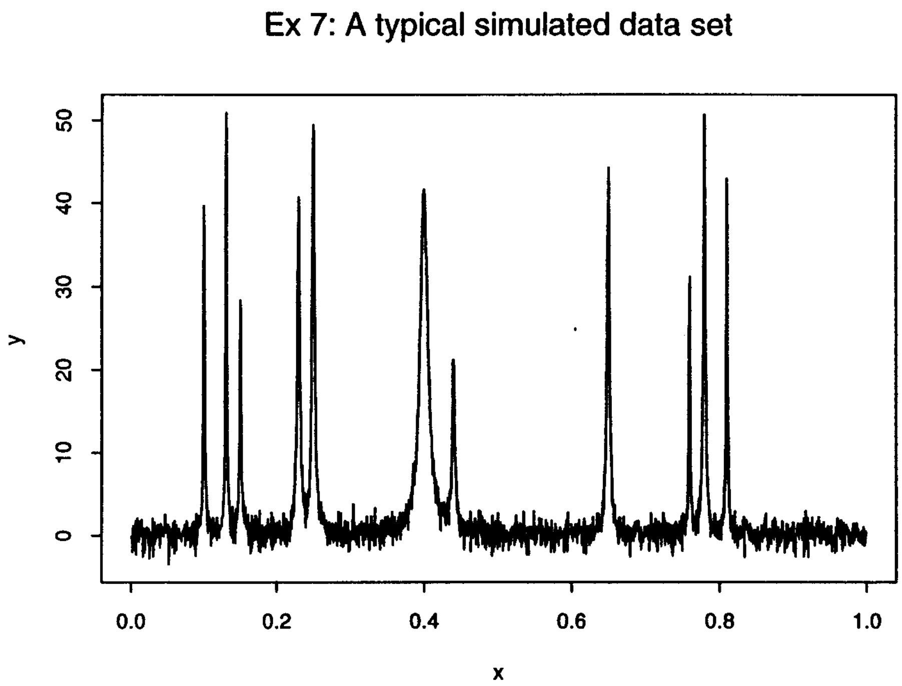
Ex 7: A typical simulated data set

Figure 7.a

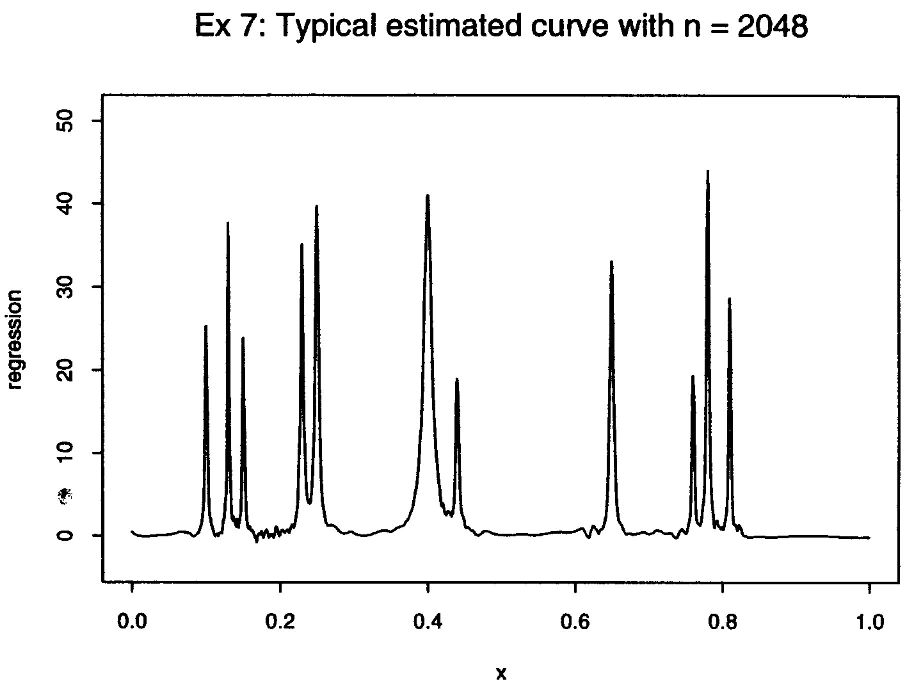
Ex 7: Typical estimated curve with n = 2048

Figure 7.b

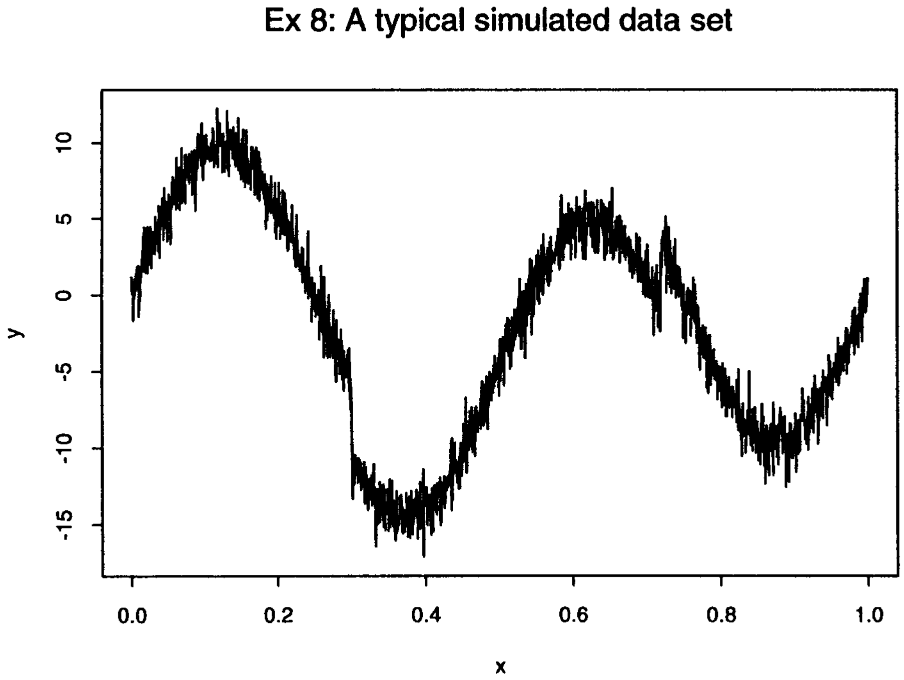
Ex 8: A typical simulated data set

Figure 8.a

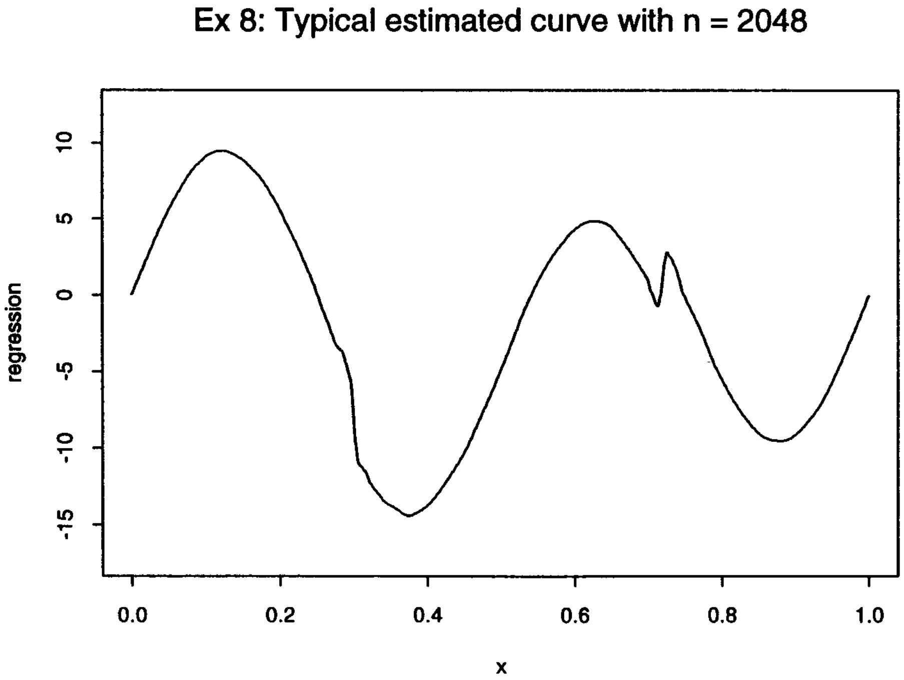
Ex 8: Typical estimated curve with n = 2048

Figure 8.b

Ex 9: Motorcycle Data

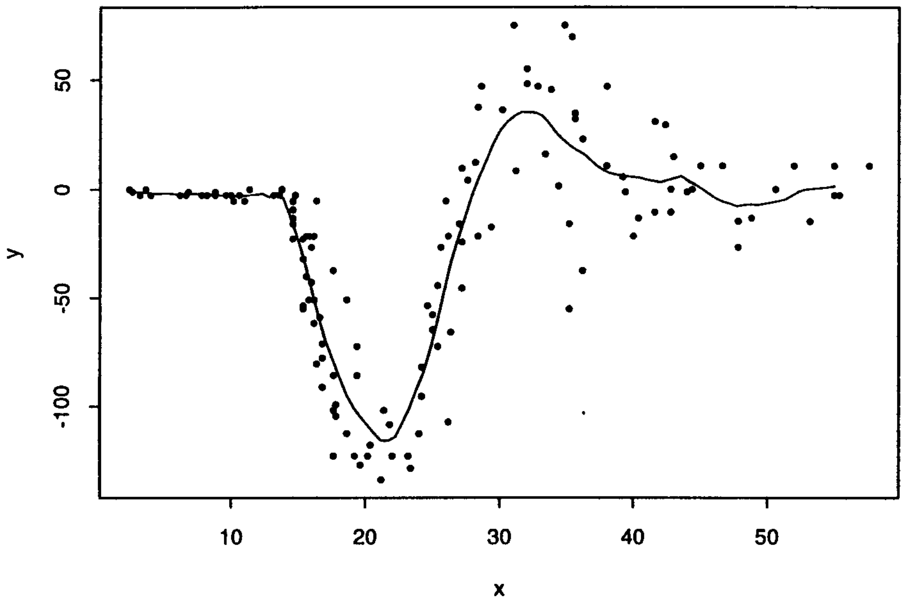

Figure 9

Ex 2: Derivative Estimation

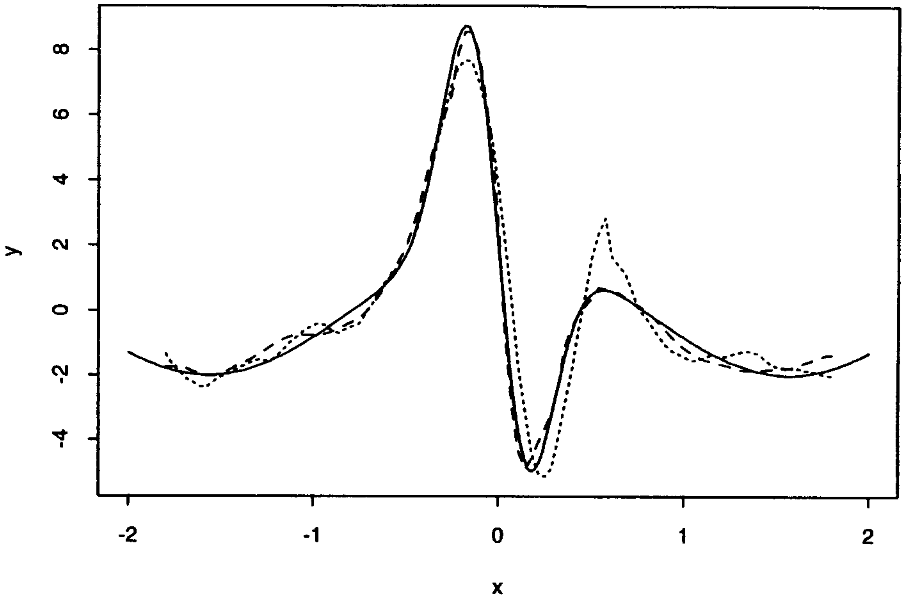

Figure 2.e

Ex 5: Derivative Estimation

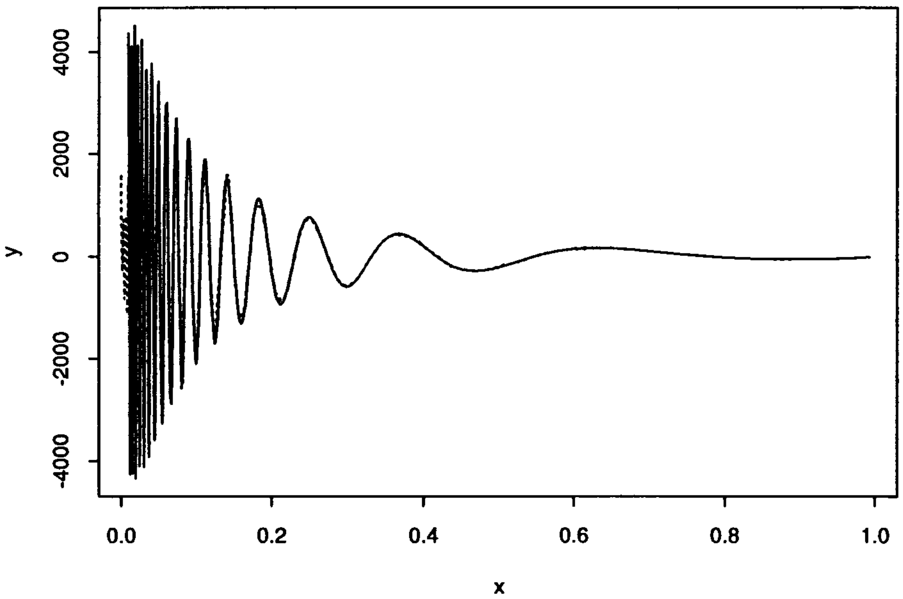

Figure 5.c