Nonparametric Regression from Independent Sources

by

Patrick D. Gerard
Experimental Statistics Unit, Box 9653
Mississippi State, MS 39762-9653

William R. Schucany
Department of Statistical Science
Southern Methodist University
Dallas, TX 75275-0332

Technical Report No. SMU/DS/TR-275

Nonparametric Regression from Independent Sources

Patrick D. GERARD
Experimental Statistics Unit, Box 9653, Mississippi State University, MS 39762-9653, USA

William R. SCHUCANY
Department of Statistical Science, Southern Methodist University, Dallas, TX 75275-0332, USA

Abstract: Three nonparametric curve estimators with information from different sources proposed by Gerard and Schucany (1995) are investigated for finite sample performance. Local bandwidth selection using an automatic procedure is adapted to local linear estimation with two independent datasets. Consistency results are provided. The previous asymptotic results are supported in two simulation trials with data-based local bandwidth selection.

Key Words:adaptive bandwidth,kernel,local linear

1.Introduction

Suppose that we have data on a response variable (y_i) and an explanatory variable (x_i). Consider estimating some unknown smooth relationship between them, often denoted by

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

Details of this model will be given in the next section. A tremendous amount has been written regarding estimation of $m(x)$ using nonparametric regression, particularly kernel regression estimators and locally weighted polynomial regression estimators. For good discussions of nonparametric regression estimators using these methods see Eubank (1988), Müller (1988), or Härdle (1990). For a recent comparison of some aspects of these two see Hastie and Loader (1993). There is a rich literature on this general problem. Relatively little has been published on data sets from independent sources.

Only a few journal articles can be found on the subject of combining independent nonparametric regression estimators. Härdle and Marron (1990) approach such a problem from a semiparametric perspective. Hart and Wehrly (1986) do not assume independence of the observations within a group and do not take full advantage of independence if it is present. Gerard and Schucany (1995) investigate the asymptotic properties of three estimators. These estimators linearly combine data from two sources, possibly with different variances, to form a nonparametric curve estimator . These may combine either kernel regression estimators or locally weighted linear regression estimators. The biases of these estimators depend upon the amount of smoothing, which is controlled by window width or bandwidth. Minimizing the usual optimality criterion of asymptotic mean squared error (AMSE) requires a balancing of bias^2 and variance. Both of these depend on the bandwidth.

The first of these estimators, $\hat{m}_N(t)$, follows the naive approach of disregarding the fact that the data come from two sources and proceeding for example with locally weighted

linear regression, as though one large dataset were available. That such an approach will be employed becomes increasingly likely as nonparametric curve estimators find their way into widely used computer packages. This naive estimator is equivalent, in terms of AMSE, to a linear combination of two nonparametric regression estimators, each employing a common bandwidth. The variances are not involved in the weights so it is not possible for this estimator to reduce the influence of a more variable data set. Thus it performs poorly when one group has a much larger variance.

The second of these estimators, $\hat{m}_O(t)$, is also a linear combination of estimators from each group. However, the two asymptotically optimal bandwidths for the two individual estimators are used. The weights are then obtained by minimizing the associated AMSE of that combination.

The third estimator, $\hat{m}_E(t)$, results from solving the more general problem of minimizing the AMSE of a linear combination of nonparametric regression estimators simultaneously with respect to the weighting factors and the two bandwidths. Equal bandwidths and a weighting factor proportional to the inverse of the variance divided by the product of sample size and design density yield a local minimum AMSE.

The AMSE of $\hat{m}_E(t)$ is never greater than that of $\hat{m}_N(t)$. Equality occurs when the ratio of the respective variance divided by the product of sample size and design density is unity. In particular, this includes the special balanced case of equal variances and an equal number of design points from the same design density for each group. Comparison of $\hat{m}_E(t)$ with $\hat{m}_O(t)$ is facilitated by the fact that the ratio of their respective AMSE's is a function of the ratio specified above. Detailed comparisons of the three AMSE's can be found in Gerard and Schucany (1995).

Before any of these estimators can be used in practice, a bandwidth must be chosen. Global bandwidths, (the same bandwidth at each estimation point), are typically chosen using some variant of cross validation or plug-in estimation (Hall *et. al.* , 1991; Gasser *et. al.* ,1991). Curve estimation with local bandwidths, (different bandwidths at each point), has been shown (Müller and Stadtmüller, 1987) to be superior in terms of AMSE to estimation with global bandwidths. Brockman *et. al.* (1993) investigate an iterative plug-in bandwidth method that requires a fixed number of iterations. Schucany (1995) proposes a plug-in method that employs derivative estimates obtained from least squares smoothing of bias estimates at a fixed grid of bandwidths. This method for kernels is adapted in Section 3 for locally weighted linear regression estimation. A further adaptation is proposed for boundary cases. Local bandwidth selection procedures are developed using this methodology for the three estimators combining local linear fits. The results of two simulation studies using these estimators are summarized in Section 4. They are consistent with the asymptotic results of Gerard and Schucany (1995).

2. Methodology

2.1 Background

Nonparametric regression methods such as kernel regression and locally weighted linear regression have become a reasonable choice for curve estimation without specifying a

functional form. The responses, y_i, are assumed to be related to the explanatory variables, x_i, by

$$y_i = m(x_i) + \varepsilon_i, \qquad i = 1,\ldots,n, \tag{2.1}$$

with ε_i being independent, identically distributed random variables having zero mean and constant variance σ^2. The x_i are fixed values satisfying $x_i = F^{-1}\left(i/(n+1)\right)$, where $F(\cdot)$ is an absolutely continuous cumulative distribution function with corresponding density $f(\cdot)$, known as the design density. Usually a certain amount of smoothness is assumed for $m(\cdot)$.

Locally weighted polynomial regression estimators (Fan,1992) have gained popularity to some extent because of their improved performance in boundary regions (near the edges of their available data) compared to kernel estimators. The "local linear" estimators are

$$\hat{m}_{LL}(t;h) = \frac{\sum_{i=1}^{n} K\left(\frac{z_i}{h}\right) z_i^2 \sum_{i=1}^{n} K\left(\frac{z_i}{h}\right) y_i - \sum_{i=1}^{n} K\left(\frac{z_i}{h}\right) z_i \sum_{i=1}^{n} K\left(\frac{z_i}{h}\right) z_i y_i}{\sum_{i=1}^{n} K\left(\frac{z_i}{h}\right) z_i^2 \sum_{i=1}^{n} K\left(\frac{z_i}{h}\right) - \left(\sum_{i=1}^{n} K\left(\frac{z_i}{h}\right) z_i\right)^2}, \tag{2.2}$$

where $z_i = x_i - t$. The weight function, $K(\cdot)$, is typically a second order kernel function supported on [-1,1] and h is a bandwidth governing the smoothness of the estimator. Larger values of h produce smoother curves, but the bias in $\hat{m}(t)$ is greater as a result.

Performance of these estimators is typically gauged by AMSE. For interior estimation ($h < t < 1-h$) the AMSE is (Jones et. al.,1994),

$$AMSE\left[\hat{m}_{LL}(t;h)\right] = \left[\frac{1}{2} k_2 m''(t) h^2\right]^2 + \frac{\sigma^2 Q}{nhf(t)}, \tag{2.3}$$

where for second-order kernels, $k_2 = \int_{-1}^{1} u^2 K(u)\,du$ and $Q = \int_{-1}^{1} K^2(u)\,du$. The only assumptions required are that $n \to \infty$ and $h \to 0$ such that $nh \to \infty$ and some continuity of $m''(t)$. Consequently, our findings in this paper for local linear regression with weight function $K(\cdot)$ apply immediately for kernel regression with that same kernel.

When data are available from two sources, it may be reasonable to assume that the $j^{\underline{th}}$ observation in the $i^{\underline{th}}$ group, y_{ij}, is related to the corresponding explanatory variable , x_{ij}, through the same smooth mean function, i.e.,

$$y_{ij} = m(x_{ij}) + \varepsilon_{ij}, \qquad i = 1,2, j = 1,\ldots,n_i, \tag{2.4}$$

where the ε_{ij} are independent and identically distributed within the ith group with zero mean and variance σ_i^2. Again, the x_{ij} are fixed variables satisfying $x_{ij} = F_i^{-1}(j/(n_i+1))$, where $F_i(\cdot)$ is an absolutely continuous cumulative distribution function with corresponding density $f_i(\cdot)$. Different designs and different variances provide a reasonably flexible model. A practical application of this model might involve blood cholesterol as a function of body fat content. If data are available on two groups of subjects from two laboratories with equipment that measures the response variable with different precision, then this model is appropriate.

2.2 Estimators that Combine Nonparametric Regressions

The naive estimator, $\hat{m}_N(t)$, performs locally weighted linear regression as though only one "pooled" dataset were available. If the usual assumptions of $n_1 \to \infty$, $n_2 \to \infty$, and $h \to 0$ such that $n_1 h \to \infty$ and $n_2 h \to \infty$ are supplemented with $n_1/n_2 \to r$ $(0 < r < \infty)$, then using standard integral approximations and Taylor series expansions the expressions for the asymptotic bias and variance of this estimator, $\hat{m}_N(t)$, are

$$Bias[\hat{m}_N(t)] = \frac{1}{2} k_2 m''(t) h^2 + o(h^2)$$

and

$$Var[\hat{m}_N(t)] = \frac{Q}{h}\left[\frac{\sigma_1^2 n_1 f_1(t) + \sigma_2^2 n_2 f_2(t)}{(n_1 f_1(t) + n_2 f_2(t))^2}\right] + o\left(\frac{1}{n_1 h}\right).$$

Details of these derivations may be found in Gerard (1993).

Introducing notation for an equivalent variance, namely

$$\sigma_N^2 = \frac{\sigma_1^2 n_1 f_1(t) + \sigma_2^2 n_2 f_2(t)}{(n_1 f_1(t) + n_2 f_2(t))^2},$$

the asymptotically optimal bandwidth, h_N has the familiar form

$$h_N = \left[\frac{\sigma_N^2 Q}{(k_2 m''(t))^2}\right]^{1/5}.$$

This estimator is easily seen to be equivalent, in terms of AMSE, to

$$\hat{m}^*(t;h) = \frac{n_1 f_1(t)}{n_1 f_1(t) + n_2 f_2(t)} \hat{m}_1(t;h) + \frac{n_2 f_2(t)}{n_1 f_1(t) + n_2 f_2(t)} \hat{m}_2(t;h),$$

where $\hat{m}_i(t;h)$ is the locally weighted linear regression estimator (2.2) from the $i^{\underline{th}}$ group. As can be seen from the expression for $\hat{m}^*(t;h)$, this estimator essentially weights the individual estimators by the amount of data available to each and does not involve the variances σ_i^2.

A seemingly more reasonable method of estimation entails linearly combining estimators from each group, each of which employs the asymptotically optimal bandwidth for that respective group. The combination is selected that minimizes the AMSE of the resulting estimator. That is, consider the class of estimators

$$\hat{m}_O(t;c) = c\hat{m}_1(t;hopt_1) + (1-c)\hat{m}_2(t;hopt_2),$$

where $hopt_i$ is the asymptotically optimal bandwidth for the $i^{\underline{th}}$ group, derived separately from (2.3),

$$hopt_i = \left[\frac{\sigma_i^2 Q}{(k_2 m''(t))^2 n_i f_i(t)}\right]^{1/5}, \quad i=1,2. \tag{2.5}$$

It follows that the value of c that minimizes $AMSE\left[\hat{m}_O(t;c)\right]$ is easily obtained by differentiation to be

$$c_O = \frac{5-k^{2/5}}{5k^{4/5}-2k^{2/5}+5}, \text{ where } k = \frac{\sigma_1^2}{n_1 f_1(t)} \Big/ \frac{\sigma_2^2}{n_2 f_2(t)}. \tag{2.6}$$

This is the ratio that was mentioned in the introduction. Use of this value of c yields $\hat{m}_O(t)=\hat{m}_O(t;c_O)$.

This second estimator was derived by minimizing the AMSE of a linear combination with each of the bandwidths fixed at their asymptotically optimal values. If, instead, the bandwidths as well as the weighting factor are allowed to vary, then a more general minimization problem arises. Gerard and Schucany (1995) showed that using

$$c = c_{opt} = \frac{\dfrac{\sigma_2^2}{n_2 f_2(t)}}{\dfrac{\sigma_1^2}{n_1 f_1(t)} + \dfrac{\sigma_2^2}{n_2 f_2(t)}} \tag{2.7}$$

and

$$h_1 = h_2 = h_E = \left[\frac{\dfrac{\sigma_1^2}{n_1 f_1(t)} \dfrac{\sigma_2^2}{n_2 f_2(t)} Q}{\left(\dfrac{\sigma_1^2}{n_1 f_1(t)} + \dfrac{\sigma_2^2}{n_2 f_2(t)} \right) (k_2 m''(t))^2} \right]^{1/5}, \tag{2.8}$$

the estimator

$$\hat{m}_E(t) = \frac{\dfrac{\sigma_2^2}{n_2 f_2(t)}}{\dfrac{\sigma_1^2}{n_1 f_1(t)} + \dfrac{\sigma_2^2}{n_2 f_2(t)}} \hat{m}_1(t;h_E) + \frac{\dfrac{\sigma_1^2}{n_1 f_1(t)}}{\dfrac{\sigma_1^2}{n_1 f_1(t)} + \dfrac{\sigma_2^2}{n_2 f_2(t)}} \hat{m}_2(t;h_E),$$

yields a local minimum mean squared error among estimates of the form

$$\hat{m}(t;c,h_1,h_2) = c\hat{m}_1(t;h_1) + (1-c)\hat{m}_2(t;h_2).$$

Additionally, $\hat{m}_E(t)$ results in a global minimum asymptotic AMSE among estimators of this form with the bandwidths constrained to be equal.

Asymptotic relative efficiency (ARE) comparisons of these estimators, made through ratios of minimum AMSE's, are facilitated by the fact that those ratios depend only on k. Only one estimator is uniformly better than another. The "naive" is virtually dominated by the "equal bandwidth" approach, i.e. the minimized AMSE of $\hat{m}_N(t)$ is as least as large as that of $\hat{m}_E(t)$ for every k. See Gerard and Schucany (1995) for ARE's that favor $\hat{m}_E$ to $\hat{m}_O$ as long as $1/161.08 < k < 161.08$.

In this paper we show that finite sample practical implementations have relative efficiencies that agree with large sample theory.

3. Bandwidth Selection

3.1 Adaptive Bandwidth (AB) Method for Local Linear Regression

Consider the model described in (2.1). For interior estimation, the asymptotic mean squared error of a locally weighted linear regression estimator (2.3) is

$$\begin{aligned} AMSE\left[\hat{m}_{LL}(t;h)\right] &= \left[\frac{1}{2} k_2 m''(t) h^2\right]^2 + \frac{\sigma^2 Q}{nf(t)h} \\ &= \frac{1}{4}\left[k_2 m''(t)\right]^2 h^4 + \frac{\sigma^2 Q}{nf(t)h} \end{aligned} \tag{3.1}$$

$$= Bh^4 + \frac{\sigma^2 A}{nh},$$

where $B = \frac{1}{4}\left[k_2 m''(t)\right]^2$, $A = \frac{Q}{f(t)}$, and k_2 and Q are as in (2.3). The optimal bandwidth is

$$h_{opt} = \left[\frac{\sigma^2 A}{4nB}\right]^{1/5}.$$

Since A contains known quantities, only estimates of σ^2 and B are required to estimate h_{opt}. Numerous $\sqrt{n}$-consistent estimators of σ^2 exist; see for example Gasser, Sroka, and Jennen-Steinmetz (1986). Schucany (1995) proposed an estimator for B motivated by the form of the desired asymptotic bias,

$$bias^2\left[\hat{m}_{LL}(t;h)\right] = Bh^4 + o\left(h^4\right). \tag{3.2}$$

If the bias were known for a number of bandwidths, $h_1,....,h_q$, then the relationship between the squared bias and the bandwidth could be written as the regression model

$$b_j^2 = bias^2\left[\hat{m}_{LL}(t;h_j)\right] = Bh_j^4 + o(h_j^4), \quad j = 1....q,$$

where the dependent variable is the squared bias, the independent variable is the bandwidth, and the remainder term represents the model error. Since the bias is not known, it must be estimated. One consistent estimator employing differences of locally weighted polynomial estimators is

$$\hat{b}_j = \hat{m}_{LL}(t;h_j) - \hat{m}_{LQ}(t;h_j), \quad j = 1...q, \tag{3.3}$$

where $\hat{m}_{LQ}(t;h)$ is a locally weighted quadratic regression estimator.

The estimator of B, obtained using least squares, is

$$\hat{B} = \frac{\sum_{j=1}^{q} h_j^4 \hat{b}_j^2}{\sum_{j=1}^{q} h_j^8}. \tag{3.4}$$

Asymptotic properties of $\hat{B}$ and the resulting plug-in estimator of h_{opt} are given in the following theorem. The theorems require that $m''(t)$ satisfy a Lipschitz condition of order γ. A function $g(\cdot)$ satisfies a Lipschitz condition of order γ over the range [0,1] if for all $a,b \in$ [0,1],

$$|g(a)-g(b)| \le M|a-b|^{\gamma}, \quad |M| < \infty, \quad 0 < \gamma \le 1.$$

Proofs of this and subsequent theorems are relegated to the Appendix.

Theorem 1

Assume that $m''(t)$ satisfies a Lipschitz condition of order γ and that $f'(t)$ is finite for all $t \in [0,1]$. Let $\hat{b}_j$ in (3.3) be an estimator of $bias\left[\hat{m}_{LL}(t;h_j)\right]$. If the grid $h_j = C_j n^{-\rho}$ with $\rho>0$ and $|C_j| < \infty, j=1,\ldots,q$, then

$$i) \qquad \hat{B} = \frac{\sum_{j=1}^{q} h_j^4 \hat{b}_j^2}{\sum_{j=1}^{q} h_j^8} = B + O\left(n^{-\rho\gamma}\right) + O_p\left(n^{-(1-5\rho)/2}\right),$$

and the AB method is consistent in that

$$ii) \qquad \hat{h} = \left(\frac{\hat{\sigma}^2 A}{4n\hat{B}}\right)^{1/5} = h_{opt}\left(1 + O_p\left(n^{-(1-5\rho)/2}\right) + O\left(n^{-\gamma\rho}\right)\right),$$

$$\text{if } \hat{\sigma}^2 = \sigma^2 + O_p\left(n^{-1/2}\right).$$

In the next section, the AB method is adapted for use in combining information from different sources with $\hat{m}_E(t)$.

3.2 Estimation of Common Bandwidth, h_E

The asymptotically optimal bandwidth of $\hat{m}_E(t)$, given in (2.8), also depends on B by substituting $4B = \left(k_2 m''(t)\right)^2$, which follows from (3.1) and (3.2). However, in this situation, there are two sources for the bias estimate since the asymptotic biases of $\hat{m}_1(t;h)$ and $\hat{m}_2(t;h)$ are equal for the same bandwidth and weight function. To be consistent with the premise that the mean function is the same for each group as in (2.4), a combined estimate of bias is used to estimate B. If a grid of trial bandwidths is used to estimate the bias for each group, then a linear combination of the bias estimates could be used to estimate B as in the previous section. The optimal linear combination of bias estimates is exactly the same as for combining $\hat{m}_1$ and $\hat{m}_2$.

Theorem 2

Assume that the model (2.4) holds. Let $n_1 \to \infty$, $n_2 \to \infty$, and $h \to 0$ such that $n_1/n_2 \to r$ $(0<r<\infty)$ and $n_2 h \to \infty$. Assume also that the conditions of Theorem 1 hold. If $h=O(n^{-\rho})$ and $\rho > 1/(5+2\gamma)$, then the minimum AMSE value of c for the linear combination $\hat{b} = c\hat{b}_1 + (1-c)\hat{b}_2$ is the same c_{opt} given in (2.7).

Inserting suitable variance estimates to obtain an estimate of this c_{opt} yields a combined estimate of bias, $\hat{b}_c$. These for a grid of bandwidths can then be used to estimate B as before in (3.4). The next theorem gives some asymptotic properties of $\hat{B}$ and the resulting plug-in estimator of h_E.

Theorem 3

Assume that the conditions of Theorem 3 hold. Define $\hat{b}_{cj}$ as the estimated optimal combination,

$$\hat{b}_{cj} = \frac{\dfrac{\hat{\sigma}_2^2}{n_2 f_2(t)}}{\dfrac{\hat{\sigma}_1^2}{n_1 f_1(t)} + \dfrac{\hat{\sigma}_2^2}{n_2 f_2(t)}} \hat{b}_{1j} + \frac{\dfrac{\hat{\sigma}_1^2}{n_1 f_1(t)}}{\dfrac{\hat{\sigma}_1^2}{n_1 f_1(t)} + \dfrac{\hat{\sigma}_2^2}{n_2 f_2(t)}} \hat{b}_{2j}, \quad j = 1,2,\ldots q.$$

If $\hat{\sigma}_i^2 = \sigma_i^2 + O_p\left(n_i^{-1/2}\right)$, for i=1,2, then

$$i) \qquad \hat{B} = \frac{\sum_{j=1}^{q} \hat{b}_{cj}^2 h_j^4}{\sum_{j=1}^{q} h_j^8} = B + O\left(n_1^{-\rho\gamma}\right) + O_p\left(n_1^{-(1-5\rho)/2}\right)$$

and

$$ii) \qquad \hat{h}_E = \left[\frac{\hat{\sigma}_E^2 A}{4\hat{B}}\right]^{1/5} = h_E\left[1 + O_p\left(n_1^{-(1-5\rho)/2}\right) + O\left(n_1^{-\rho\gamma}\right)\right],$$

$$\text{if } \hat{\sigma}_E^2 = \frac{\dfrac{\hat{\sigma}_1^2}{n_1 f_1(t)} \dfrac{\hat{\sigma}_2^2}{n_2 f_2(t)}}{\dfrac{\hat{\sigma}_1^2}{n_1 f_1(t)} + \dfrac{\hat{\sigma}_2^2}{n_2 f_2(t)}} \text{ and } h_E \text{ is given in (2.8).}$$

3.3 Boundary Modifications

For interior estimation, $AMSE\left[\hat{m}_{LL}(t;h)\right]$ is given by (2.3). However, when the point of estimation falls in the left boundary ($0<t<h$), Fan (1992) gives

$$AMSE\left[\hat{m}_{LL}(t;h)\right] = \frac{1}{4}\left(\beta m''(0)\right)^2 h^4 + \frac{\alpha\sigma^2}{nhf(0)}. \tag{3.5}$$

Here taking $S_{p,d} = \int_{-1}^{d} K(u)u^p du$, ($p$=0,1,2,3) and $d=t/h<1$, the new factors are

$$\beta = \frac{S_{2,d}^2 - S_{1,d}S_{3,d}}{S_{2,d}S_{0,d} - S_{1,d}^2} \text{ and } \alpha = \frac{\int_{-1}^{d}\left(S_{2,d} - uS_{1,d}\right)^2 K^2(u)du}{\left(S_{2,d}S_{0,d} - S_{1,d}^2\right)^2}.$$

Strictly speaking, α and β are functions of h due to their dependence on d. However, these asymptotic expansions are derived assuming that d remains constant as $h\to 0$. To do so the point of estimation *moves* toward the endpoint of the estimation space as $h\to 0$ so that it remains in the boundary region. Analogous expressions hold for the right boundary.

In practice, the point of estimation is constant and α and β must be viewed as functions of h denoted by $\alpha(h)$ and $\beta(h)$. Finding the value for h that minimizes (3.4) is not the simple exercise in differentiation encountered in interior situations. Even if the point of estimation remains constant, a good approximation for the boundary is

$$\begin{aligned} AMSE\left[\hat{m}_{LL}(t;h)\right] &= \frac{1}{4}\left(\beta(h)m''(t)\right)^2 h^4 + \frac{\alpha(h)\sigma^2}{nhf(t)} \\ &= B_B\left(\beta(h)\right)^2 h^4 + \frac{\alpha(h)A_B\sigma^2}{nh}, \end{aligned}$$

where $B_B = \frac{1}{4}\left(m''(t)\right)^2$ and $A_B=1/f(t)$. The algorithm for local bandwidth selection is modified as follows. As before, A_B contains known quantities and B_B is estimated from smoothed bias estimates from a grid of bandwidths. Starting values of $\beta(h)=k_2$ and $A_B\alpha(h)=Q/f(t)$ are plugged into (2.5) to estimate a bandwidth. If the estimated bandwidth results in boundary conditions, then there is a search for the bandwidth that "minimizes" the

estimated asymptotic mean squared error, $\hat{B}_B\beta^2(h)h^4 + \frac{A_B\alpha(h)\hat{\sigma}^2}{nh}$. This expression may have more than one local minimum. When this occurs, the smallest bandwidth producing a local minimum is chosen. This is consistent with the local philosophy underlying nonparametric regression estimators. It also may prevent problems caused by the lack of agreement between actual and asymptotic conditions, especially for large bandwidths. There is an analogous modification to be made for the bandwidth selection procedure for h_E.

In the next section, we compare the finite sample performance of our three curve estimators in two simulation studies. The first study includes all three estimators and the second concentrates solely on the two more reasonable estimators, $\hat{m}_E(t)$ and $\hat{m}_O(t)$.

4. Simulation Results

4.1 Pilot Simulation Study

In this section, finite sample properties of $\hat{m}_N(t)$, $\hat{m}_O(t)$, and $\hat{m}_E(t)$ will be compared by estimating integrated mean squared error with

$$\widehat{IMSE}[\hat{m}] = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}(x_i) - m(x_i)\right)^2, \tag{4.1}$$

where $\hat{m}$ is either $\hat{m}_N$, $\hat{m}_O$, or $\hat{m}_E$. These summaries will be averaged over independent replicates.

The purpose of this study is to investigate the effect of the ratio, k, defined in (2.6) on the integrated mean squared error of these three estimators. A function with constant second derivative, $m(x)=x(1-x)$, is used. For each of the two groups, observations were generated at 100 equally spaced design points. Since this reduces to uniform design densities for each group with equal sample sizes, the value of k is simply the ratio of variances, $k = \sigma_1^2/\sigma_2^2$. The values of k studied are 1, 10, 50, 150, and 200. For k=1, the standard deviation for each group equals 25% of the range of the mean function. For the other values of k, group 1 has the same standard deviation and the standard deviation of group 2 is reduced to give the desired value of k. An example of a pair of realizations with k=100 is given in Figure 1.

Local bandwidths were used for each of the three estimators. The methods described in the previous section produced local bandwidths for $\hat{m}_E$. For $\hat{m}_O$, individual bandwidths were estimated for each group from (2.5) using the same $\hat{B}$ obtained from the combined bias estimates described in Theorem 3. A general expression for c_O to be used in $\hat{m}_O(t)$ is

$$c_O = \frac{v_2 - \left(b_1 b_2 - b_2^2\right)}{(b_1 - b_2)^2 + v_1 + v_2},$$

where b_1, v_1, b_2, and v_2 are the bias and variances of the locally weighted linear regression estimators for the two groups. Using $\hat{h}_1$, $\hat{h}_2$, $\hat{\sigma}_1^2$, and $\hat{\sigma}_2^2$, the general expression for an estimate of c_o is

$$\hat{c}_O = \frac{\dfrac{A_{B2}\alpha(\hat{h}_2)\hat{\sigma}_2^2}{n_2\hat{h}_2} - \left(\hat{B}_B\beta(\hat{h}_1)\beta(\hat{h}_2)\hat{h}_1^2\hat{h}_2^2 - \hat{B}_B\hat{h}_2^4\left(\beta(\hat{h}_2)\right)^2\right)}{\hat{B}_B\left(\beta(\hat{h}_1)\hat{h}_1^2 - \beta(\hat{h}_2)\hat{h}_2^2\right)^2 + \dfrac{A_{B2}\alpha(\hat{h}_2)\hat{\sigma}_2^2}{n_2\hat{h}_2} + \dfrac{A_{B1}\alpha(\hat{h}_1)\hat{\sigma}_1^2}{n_1\hat{h}_1}},$$

where $A_{Bi}=1/f_i(t)$. The resulting estimator is thus

$$\hat{m}_O(t) = \hat{c}_O\hat{m}_1(t;\hat{h}_1) + (1-\hat{c}_O)\hat{m}_2(t;\hat{h}_2).$$

For $\hat{m}_N$, the data were combined into one group and the AB method employed directly.

The grid of fixed bandwidths must be specified for local bandwidth selection. After some empirical investigation,

$$h_j = (j/7)(n_1+n_2)^{-1/7}, \quad j = 1,\ldots,7,$$

was found to be a reasonable choice, where n_1 and n_2 are the sample sizes for the first and second groups, respectively.

Independently, for each of the values of k, 100 replicates were generated. The estimates of integrated mean squared error are summarized in Table 1 for six values of k. Significance tests are based on paired differences between values of (4.1) for each replicate.

Table 1. Monte Carlo Estimates of Integrated Mean Square Error
Averages of (4.1) are multiplied by 10^5 as well as their estimated standard errors in parentheses

k	$\hat{m}_N$	$\hat{m}_O$	$\hat{m}_E$
1	13.583(.769)	13.882(.796)	13.556(.757)
10	7.403(.332)	3.185(.140)	3.067(.135)
50	7.015(.293)	0.882(.035)	0.865(.035)
100	7.349(.353)	0.486(.020)	0.489(.021)
150	6.707(.315)	0.340(.013)	0.340(.013)
200	6.951(.380)	0.243(.009)	0.245(.009)

The results tend to support the asymptotic findings of Gerard and Schucany (1995). For k=1, there is little difference between $\hat{m}_N(t)$ and $\hat{m}_E(t)$ (p=.6181 for signed-rank test).

Each performs better than $\hat{m}_O(t)$, but the comparison with $\hat{m}_N$ not statistically significant (p=.1824) while the comparison with $\hat{m}_E$ is (p=.0304). As k increases, the performance of $\hat{m}_N$ relative to the other two estimators deteriorates (p<.05 in all cases), also in agreement with the asymptotic results. For k=10 and k=50, $\hat{m}_E$ performed better than $\hat{m}_O$ (p=.0001 and p=.0024, respectively). For k larger than 50, there is no statistically significant differences between the two estimators. In the next section, a larger simulation trial involving only $\hat{m}_O(t)$ and $\hat{m}_E(t)$ is discussed.

4.2 Simulation Comparison of Two Contenders

The restriction in the pilot study of equal numbers and equal spacing is relaxed here. The same mean function is employed, but two different design densities, two samples sizes, and two different standard deviations are employed. The seven combinations studied are listed in Table 2, where Δ is the range of m (max-min=.25). A complete factorial arrangement of the factors was not undertaken because the number of factor-level combinations would be prohibitive. Furthermore, due to the symmetry of this problem, a complete factorial arrangement would exactly duplicate some factor-level combinations, simply switching group designations.

Table 2. Simulation Study Cases

	Group 1		Group 2	
Case	Sample size (n_1)	Standard deviation (σ_1)	Sample size (n_2)	Standard deviation (σ_2)
1	100	.25 Δ	100	.10 Δ
2	200	.25 Δ	200	.10 Δ
3	100	.25 Δ	200	.10 Δ
4	100	.10 Δ	100	.10 Δ
5	100	.25 Δ	100	.25 Δ
6	200	.10 Δ	200	.10 Δ
7	100	.25 Δ	200	.25 Δ

Note that use of these combinations results in values of k ranging from 1 (Cases 4 and 5) to 12.5 (Case 3) and encompasses situations that reasonably might be seen in practice.

For each case described in Table 2, 100 replicates were generated for each of two design densities. The two densities are $f_1(x)=1$ and $f_2(x)=e^{-x}/(1-e^{-1})$, both supported on the interval (0,1). Note that $f_2(x)$ is a truncated exponential distribution. The same grid of bandwidths is used as in the pilot study. For each of the 100 replications, $\hat{m}_E$ and $\hat{m}_O$ were evaluated and another estimate of integrated mean squared error,

$$\hat{IMSE}[\hat{m}] = \frac{1}{50}\sum_{i=1}^{50}\left(\hat{m}(xe_i) - m(xe_i)\right)^2 ,$$

(4.2)

is calculated, where $xe_i = i/51$ are arbitrary fixed evaluation points rather than actual design points. The results of the simulation study are summarized in Table 3.

Table 3. Simulation Study Summary Statistics
Averages of (4.2) are multiplied by 10^5 as well as their estimated standard errors in parentheses

Case	Design Density	$\hat{m}_O$	$\hat{m}_E$
1	f_1	3.98(.200)	3.85(.192)
1	f_2	4.24(.213)	4.11(.201)
2	f_1	2.22(.104)	2.17(.100)
2	f_2	2.71(.146)	2.66(.141)
3	f_1	2.48(.129)	2.40(.122)
3	f_2	2.46(.122)	2.40(.121)
4	f_1	2.91(.133)	2.78(.121)
4	f_2	2.96(.150)	2.82(.137)
5	f_1	12.8(.761)	12.4(.710)
5	f_2	15.2(.922)	14.7(.890)
6	f_1	1.47(.071)	1.42(.066)
6	f_2	1.65(.092)	1.59(.085)
7	f_1	9.72(.506)	9.32(.473)
7	f_2	9.16(.508)	8.79(.474)

This simulation study supports the claim that for most situations seen in practice, $\hat{m}_E$ is preferred over $\hat{m}_O$. In each of the fourteen comparisons made, the *IMSE* for $\hat{m}_E$ is significantly smaller than that of $\hat{m}_O$ ($p<.05$ for the Wilcoxon signed-rank test). It also appears that for this function, the uniform design density may be preferred over the truncated exponential since it yielded lower *IMSE* values in ten of the fourteen comparisons. Additional simulations with $m(x)$ being a combination of exponentials were also in generally good agreement with the theory.

5. Conclusions

The properties of the three estimators studied by Gerard and Schucany (1995) were investigated in finite samples through two simulation studies. Local bandwidth selection using a method suggested by Schucany (1995) was modified to estimate the smoothing parameters for these simulation studies. Consistency of the new bandwidth estimation procedure holds under usable conditions. The results of both simulation studies support the asymptotic results that $\hat{m}_N$ is very inefficient in cases with different variances, $\hat{m}_O$ is to be considered only in cases with vastly different variances, and $\hat{m}_E$ is preferred in most cases that would be encountered in practice. It should be reasonably straightforward to extend the methodology to several groups.

6. Appendix

In this section, details are given concerning proofs of the theorems cited in the body of the paper. The following three lemmas are stated without proof. The first lemma concerns the approximation of summations by integrals. A similar result is given in Conte and DeBoor (1980). The last two lemmas demonstrate the asymptotic equivalence of second and fourth order kernel estimators with locally weighted linear and quadratic regression estimators, respectively. Similar results may be found in Daniel (1992) and Müller (1988).

Lemma 1

Let $F(\cdot)$ be a cumulative distribution function with absolutely continuous density function, $f(x) = F'(x)$, strictly positive on [0,1]. If x_0=0, $x_i = F^{-1}(i/(n+1))$ for i=1, ...,n, and x_{n+1}=1, and $K(\cdot)$ is a kernel function supported on [-1,1] with finite first derivative, then

$$\frac{1}{nh}\sum_{i=1}^{n} x_i^j K\left(\frac{x_i - t}{h}\right) = \int_0^1 \frac{1}{h} x^j K\left(\frac{x-t}{h}\right) f(x)dx + O\left(n^{-1}\right), \text{ for j=0,1... , as } n\to\infty.$$

Lemma 2

Let w_{li} be the weights corresponding to a locally weighted linear regression estimator at t, $\hat{m}_{LL}(t;h)$. If the conditions of Lemma 1 hold, then

$$\hat{m}_{LL}(t;h) = \sum_{i=1}^{n} w_{li} y_i = \sum_{i=1}^{n} \frac{1}{f(t)nh} K\left(\frac{z_i}{h}\right)(1+o(1))y_i,$$

where z_i=x_i-t.

Lemma 3

Let w_{qi} be the weights corresponding to a locally weighted quadratic regression estimator at t. If the conditions of Lemma 1 hold, then

$$\hat{m}_{LQ}(t;h) = \sum_{i=1}^{n} w_{qi} y_i = \sum_{i=1}^{n} \frac{1}{nhf(t)} \left(\frac{K\left(\frac{z_i}{h}\right)\left(k_4 - k_2\left(\frac{z_i}{h}\right)^2\right)}{k_4 - k_2^2} \right)(1+o(1))y_i,$$

where $z_i = x_i - t$, and $k_i = \int_0^1 u^i K(u)du$, for i=2,4.

<u>Proof of Theorem 1</u>

i) Write $\hat{b} = \sum_{i=1}^{n} \left(w_{li} - w_{qi}\right) y_i$, where w_{li} and w_{qi} are the weights for the locally weighted linear and quadratic regression estimators respectively. Expanding $m(x_i)$ in each term with a Taylor series yields

$$\begin{aligned}\hat{b} &= \sum_{i=1}^{n} \left(w_{li} - w_{qi}\right)\left(m(x_i) + \varepsilon_i\right) \\ &= \sum_{i=1}^{n} \left(w_{li} - w_{qi}\right)\left(m(t) + m'(t)z_i + \frac{1}{2}m''(\xi_i)z_i^2 + \varepsilon_i\right), \quad \min(x_i, t) < \xi_i < \max(x_i, t),\end{aligned}$$

where $z_i = x_i - t$. Since $\sum_{i=1}^{n} w_{li} = \sum_{i=1}^{n} w_{qi} = 1$ and $\sum_{i=1}^{n} w_{li} z_i = \sum_{i=1}^{n} w_{qi} z_i = \sum_{i=1}^{n} w_{qi} z_i^2 = 0$, we have

$$\hat{b} = \sum_{i=1}^{n} \left(w_{li} - w_{qi}\right)\varepsilon_i + \frac{1}{2}\sum_{i=1}^{n} m''(\xi_i) w_{li} z_i^2 .$$

By the Lipschitz condition on $m''(x)$,

$$\begin{aligned} m''(\xi_i) &= m''(t) + O\left(|\xi_i - t|^{\gamma}\right) \\ &= m''(t) + O(h^{\gamma}), \end{aligned}$$

since $\min(x_i,t) < \xi_i < \max(x_i,t)$. Hence by Lemma 1,

$$\hat{b} = \sum_{i=1}^{n}\left(w_{li} - w_{qi}\right)\varepsilon_i + \frac{1}{2}m''(t)\sum_{i=1}^{n} w_{li} z_i^2 + O\left(h^{2+\gamma}\right)$$
$$= \sum_{i=1}^{n}\left(w_{li} - w_{qi}\right)\varepsilon_i + \frac{1}{2}m''(t)h^2 k_2 + O\left(h^3\right) + O\left(h^{2+\gamma}\right).$$

By the Central Limit Theorem (Müller, 1988),

$$\sum_{i=1}^{n}\left(w_{lq} - w_{qi}\right)\varepsilon_i = O_p\left((nh)^{-1/2}\right)$$

and hence

$$\hat{b} = \frac{1}{2}k_2 m''(t)h^2 + O\left(h^3\right) + O\left(h^{2+\gamma}\right) + O_p\left((nh)^{-1/2}\right)$$
$$= \frac{1}{2}k_2 m''(t)h^2 + O\left(h^{2+\gamma}\right) + O_p\left((nh)^{-1/2}\right).$$

Therefore

$$\hat{B} = \frac{\sum_{j=1}^{q}\hat{b}_j^2 h_j^4}{\sum_{j=1}^{q} h_j^8} = \frac{\sum_{j=1}^{q}\frac{\hat{b}_j^2}{h_j^4} h_j^8}{\sum_{j=1}^{q} h_j^8},$$

where $\hat{b}_j$ is the bias estimate when the $j^{\underline{th}}$ bandwidth, h_j. Substituting the expression for $\hat{b}_j$ yields

$$\frac{\hat{b}_j^2}{h_j^4} = B + O\left(h_j^{\gamma}\right) + O_p\left(\left(nh_j^5\right)^{-1/2}\right).$$

Substituting $h_j = C_j n^{-\rho}$ yields

$$\hat{B} = B + O\left(n^{-\rho\gamma}\right) + O_p\left(n^{-(1-5\rho)/2}\right).\|$$

ii) The ratio of the estimated bandwidth to the asymptotically optimal bandwidth is

$$\frac{\hat{h}}{h_{opt}} = \left[\frac{\hat{\sigma}^2}{\sigma^2}\frac{B}{\hat{B}}\right]^{1/5}$$

$$= \left[\frac{\sigma^2 + O_p\left(n^{-1/2}\right)}{\sigma^2}\frac{B}{B \mp O_p\left(n^{-(1-5\rho)/2}\right) + O\left(n^{-\gamma\rho}\right)}\right]^{1/5}$$

$$= \left[\left(1 + O_p\left(n^{-1/2}\right)\right)\left(1 + O_p\left(n^{-(1-5\rho)/2}\right) + O\left(n^{-\gamma\rho}\right)\right)\right]^{1/5}$$

$$= 1 + O_p\left(n^{-(1-5\rho)/2}\right) + O\left(n^{-\gamma\rho}\right).||$$

<u>Proof of Theorem 2</u>

From the proof of Theorem 1,

$$E\left(\hat{b}_i\right) = \frac{1}{2}k_2 m''(t)h^2 + O\left(h^{2+\gamma}\right),$$

and hence $Bias\left(\hat{b}_i\right) = O\left(h^{2+\gamma}\right)$.

Evaluation of $Var\left(\hat{b}_i\right)$ for i=1,2 yields from Lemmas 2 and 3,

$$Var\left(\hat{b}_i\right) = Var\left[\sum_{j=1}^{n_i}\frac{1}{n_i f_i(t)h}K\left(\frac{z_{ij}}{h}\right)(1+o(1))y_{ij} - \sum_{j=1}^{n_i}\frac{1}{n_i f_i(t)h}\left(\frac{K\left(\frac{z_{ij}}{h}\right)\left(k_4 - k_2\left(\frac{z_{ij}}{h}\right)^2\right)}{k_4 - k_2^2}(1+o(1))y_{ij}\right)\right]$$

$$= \frac{\sigma_i^2}{n_i f_i(t) h} \left[\int_{-1}^{1} K^2(u)du + \frac{\int_{-1}^{1} K^2(u)\left(k_4^2 - 2uk_2k_4 + u^2k_2^2\right)du}{\left(k_4 - k_2^2\right)^2} \right.$$

$$\left. -2\frac{\int_{-1}^{1} K^2(u)\left(k_4 - u^2k_2\right)du}{k_4 - k_2^2} \right] + o\left(\frac{1}{n_i h}\right),$$

from Lemma 1 and the continuity of $f_i(x)$. Thus,

$$\begin{aligned} MSE\left(\hat{b}_i\right) &= \left[Bias\left(\hat{b}_i\right)\right]^2 + Var\left(\hat{b}_i\right) \\ &= O\left(h^{4+\gamma}\right) + Var\left(\hat{b}_i\right) + o\left(\frac{1}{n_i h}\right) \\ &= Var\left(\hat{b}_i\right) + o\left(\frac{1}{n_i h}\right) \end{aligned}$$

if $h = O\left(n^{-\rho}\right)$ and $\rho > \frac{1}{5+2\gamma}$. The optimal value for c is found by differentiation.||

<u>Proof of Theorem 3</u>

i) From the proof of Theorem 1 and $n_1/n_2 \to r$,

$$\hat{b}_{ij} = \frac{1}{2}k_2 m''(t)h_j^2 + O\left(h_j^{2+\gamma}\right) + O_p\left(\left(n_i h_j\right)^{-1/2}\right), \quad i = 1,2, \quad j = 1,\ldots,q.$$

Thus,

$$\begin{aligned} \hat{b}_{cj} &= \left(c + O_p\left(n_1^{-1/2}\right)\right)\left(\frac{1}{2}k_2 m''(t)h_j^2 + O\left(h_j^{2+\gamma}\right) + O_p\left(\left(n_1 h_j\right)^{-1/2}\right)\right) \\ &\quad + \left(1 - c + O_p\left(n_2^{-1/2}\right)\right)\left(\frac{1}{2}k_2 m''(t)h_j^2 + O\left(h_j^{2+\gamma}\right) + O_p\left(\left(n_2 h_j\right)^{-1/2}\right)\right) \\ &= \frac{1}{2}k_2 m''(t)h_j^2 + O\left(h_j^{2+\gamma}\right) + O_p\left(\left(n_1 h_j\right)^{-1/2}\right). \end{aligned}$$

Following identical steps to those in the proof of Theorem 1 yields

$$\hat{B} = B + O\left(n_1^{-\rho\gamma}\right) + O_p\left(n_1^{-(1-5\rho)/2}\right).\|$$

ii) Consider the ratio

$$\frac{\hat{h}}{h_E} = \left[\frac{\dfrac{\hat{\sigma}_1^2\hat{\sigma}_2^2}{\left(\dfrac{\hat{\sigma}_1^2}{n_1 f_1(t)} + \dfrac{\hat{\sigma}_2^2}{n_2 f_2(t)}\right)\hat{B}}}{\dfrac{\sigma_1^2\sigma_2^2}{\left(\dfrac{\sigma_1^2}{n_1 f_1(t)} + \dfrac{\sigma_2^2}{n_2 f_2(t)}\right)B}}\right]^{1/5}.$$

Substituting $\hat{\sigma}_i^2 = \sigma_i^2 + O_p\left(n_i^{-1/2}\right)$ for i=1,2 and $\hat{B} = B + O\left(n_1^{-\rho\gamma}\right) + O_p\left(n_1^{-(1-5\rho)/2}\right)$ yields

$$\frac{\hat{h}}{h_E} = \left[\left(1 + O_p\left(n_1^{-1/2}\right)\right)\left(1 + O_p\left(n_1^{-(1-5\rho)/2}\right) + O\left(n_1^{-\gamma\rho}\right)\right)\right]^{1/5}$$
$$= 1 + O_p\left(n_1^{-(1-5\rho)/2}\right) + O\left(n_1^{-\gamma\rho}\right).\|$$

References

Brockman, M., Gasser, T., and Herrmann, E. (1993), Locally Adaptive Bandwidth Choices for Kernel Regression Estimators, *Journal of the American Statistical Association 88*, 1302-1309.

Conte, S.D. and DeBoor, C. (1980), *Elementary Numerical Analysis*, New York: McGraw Hill.

Daniel, D.L. (1992), *A Locally Weighted Least Squares Approach to Nonparametric Regression*, Ph.D. Dissertation, Department of Statistical Science, Southern Methodist University.

Eubank, R.L. (1988), *Spline Smoothing and Nonparametric Regression*, New York: Marcel Dekker.

Fan, J. (1992), Design-adaptive Nonparametric Regression, *The Journal of the American Statistical Association 87*, 998-1004.

Gasser, Th. and Müller, H.G. (1979), Kernel Estimation of Regression Functions, in *Smoothing Techniques for Curve Estimation* (Th. Gasser and M. Rosenblatt, eds.), 23-68. Heidelberg: Springer.

Gasser, Th., Sroka, L., and Jennen-Steinmetz (1986), Residual Variance and Residual Pattern in Nonlinear Regression, *Biometrika 73*, 123-127.

Gasser, Th., Kneip, A., and Köhler, W. (1991), A Flexible and Fast Method for Automatic Smoothing, *Journal of the American Statistical Association 86*, 643-652.

Gerard, P.D. (1993), *Combining Independent Nonparametric Regression Estimators,* Ph.D. Dissertation, Department of Statistical Science, Southern Methodist University.

Gerard, P.D. and Schucany, W.R. (1995), On Combining Nonparametric Regression Estimators, *Statistics and Probability Letters,* To Appear.

Härdle, W. and Bowman, A.W. (1988), Bootstrapping in Nonparametric Regression: Local Adaptive Smoothing and Confidence Bands, *Journal of the American Statistical Association 83*, 102-110.

Hastie, T. and Loader, C. (1993), Local Regression: Automatic Kernel Carpentry (with discussion), *Statistical Science 8*, 120-143.

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*, 263-269.

Härdle, W. (1990), *Applied Nonparametric Regression,* New York: Oxford University Press.

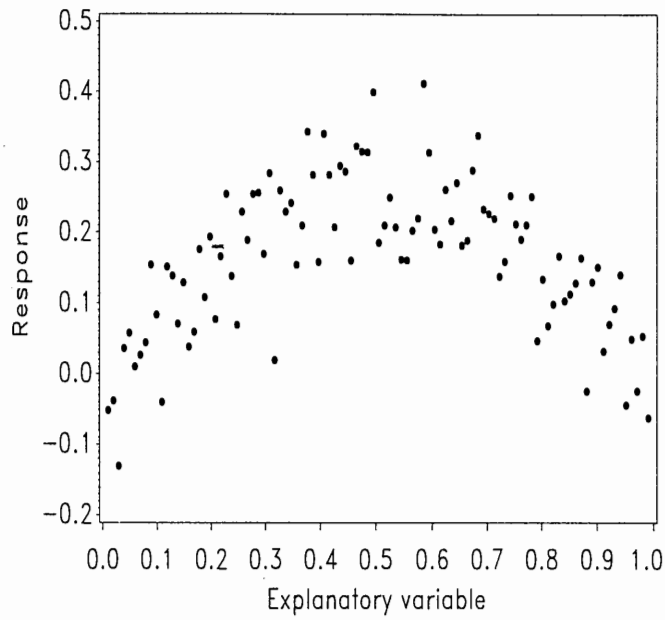
Hart, J.D. and Wehrly, T.E. (1986), Kernel Regression Estimation Using Repeated Measurements Data, *Journal of the American Statistical Association 81*, 1080-1088.

Jones,M.C., Davies, S.J., and Park,B.U. (1994), Versions of Kernel-type Regression Estimators, *Journal of the American Statistical Association 89*, 825-832.

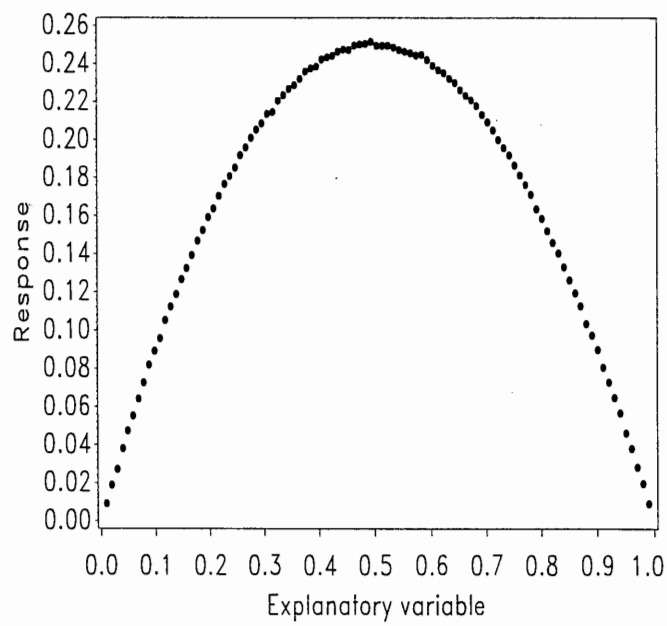
Müller, H.G. (1988), *Nonparametric Regression Analysis of Longitudinal Data*, New York: Springer-Verlag.

Müller, H.G. and Stadtmüller, U. (1987), Variable Bandwidth Kernel Estimators of Regression Curves, *The Annals of Statistics 15*, 182-210.

Schucany,W.R. (1995), Adaptive Bandwidth Choice for Kernel Regression, *Journal of the American Statistical Association* , To Appear.



a) σ_1=.25Δ



b) σ_2=.025Δ

Figure 1. Realizations with a) σ_1=.25Δ and b) σ_2=.025Δ, where Δ = range of *m*.