Local Regression: Automatic Kernel Carpentry

Trevor Hastie and Clive Loader

Abstract. A kernel smoother is an intuitive estimate of a regression function or conditional expectation; at each point x_0 the estimate of $E(Y \mid x_0)$ is a weighted mean of the sample Y_i, with observations close to x_0 receiving the largest weights. Unfortunately this simplicity has flaws. At the boundary of the predictor space, the kernel neighborhood is asymmetric and the estimate may have substantial bias. Bias can be a problem in the interior as well if the predictors are nonuniform or if the regression function has substantial curvature. These problems are particularly severe when the predictors are multidimensional.

A variety of kernel modifications have been proposed to provide approximate and asymptotic adjustment for these biases. Such methods generally place substantial restrictions on the regression problems that can be considered; in unfavorable situations, they can perform very poorly. Moreover, the necessary modifications are very difficult to implement in the multidimensional case.

Local regression smoothers fit low order polynomials in x locally at x_0, and the estimate of $f(x_0)$ is taken from the fitted polynomial at x_0. They automatically, intuitively and *simultaneously* adjust for both the biases above to the given order, and generalize naturally to the multidimensional case. They also provide natural estimates for the derivatives of f; an approach more attractive than using higher order kernel functions for the same purpose.

Key words and phrases: Boundary effects, derivative estimation, kernel, local regression, smoothing.

Trevor Hastie and Clive Loader are Members of Technical Staff at AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974.

1 Introduction

Suppose we have observations $(x_i, Y_i); i = 1, \ldots, n$, with $Y_i = f(x_i) + \epsilon_i$. Here, f is assumed to be a 'smooth' function but otherwise unknown and ϵ_i are independent errors with mean 0. The nonparametric regression problem is to estimate and find interesting structure in f.

A simple estimate proposed independently by Nadaraya (1964) and Watson (1964) is based on locally weighted averaging. Given a kernel function K, the Nadaraya-Watson (NW) estimate is

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} K(x - x_i) Y_i}{\sum_{i=1}^{n} K(x - x_i)}. \tag{1}$$

The kernel function is chosen to give most weight to observations close to x and least weight to observations far from x. Typically K is an even function specified only up to an unknown smoothing parameter h, which is selected by data based methods. Since the smoothing parameter is not a focus of this article, we suppress it in our notation for K.

Another popular estimate is the integral kernel estimate proposed by Gasser and Müller (1979). Suppose the x_i are ordered and let $s_i; i = 0, \ldots, n$ be an interpolating sequence with $s_0 \leq x_1 \leq s_1 \leq \ldots \leq x_n \leq s_n$. The Gasser-Müller (GM) estimate is defined by

$$\hat{f}(x) = \sum_{i=1}^{n} Y_i \int_{s_{i-1}}^{s_i} K(x - u) du. \tag{2}$$

Despite the simplicity, both these estimates have problems, as has been discussed by Chu and Marron (1991). One difficulty with the NW estimate is bias caused by a combination of slope in the mean function and asymmetry of observations; see Figure 1. Here, we try to estimate $f(0.6)$. Since most observations that contribute to the estimate are on the left, the estimate is slightly biased upwards. As shown in Figure 2, the use of integral kernels improves the bias. However, the weights are noisy; for example there are seven observations clustered in the interval $[0.56, 0.58]$. Although these observations all contain similar information about $f(0.6)$, they receive very different weights. This suggests the GM estimate has large variance.

Figures 1, 2 about here.

A more severe problem with both these estimates is bias in boundary regions. Suppose the x_i lie in the interval $[0,1]$, and we wish to estimate $f(0)$. In Figure 3, the mean function has substantial positive slope near 0, and hence the local average has substantial positive bias. With the GM estimate and the usual choice $s_0 = x_1, s_n = x_n$, the weights do not add to 1 at boundary points, and so the estimate can perform very poorly.

Figure 3 about here.

An alternative method of smoothing, locally weighted regression, appeared in the statistical literature in Stone (1977) and Cleveland (1979). For each point of interest x, $f(x)$ is estimated using a weighted least squares regression, with weights assigned to observations as in (1). Formally, local regression estimates can be expressed as

$$\hat{f}(x) = b(x)^T (B^T W(x) B)^{-1} B^T W(x) Y, \tag{3}$$

where $b(x)$ is an expansion of x into a basis of polynomials, B is the matrix of evaluations of b at the sample x_is, and $W(x)$ is the diagonal weight matrix implicit in (1), with

$$W_i(x) = K(x - x_i).$$

Clearly (1) is a special case of (3) with $b(x) = 1$. Implicit in (3) is the assumption that the inverse exists; for one dimensional predictors this amounts to assuming that there are at least as many unique values of x_i in the support of $K(x - \cdot)$ as there are basis functions in $b(x)$.

The local regression estimate is linear in the responses Y_j:

$$\hat{f}(x) = \sum_{j=1}^{n} l_j(x) Y_j = \langle l(x), Y \rangle , \tag{4}$$

where

$$l_j(x) = b(x)^T (B^T W(x) B)^{-1} b(x_j) W_j(x).$$

We call the weights $l_j(x)$ the *effective kernel* at x. Effective kernels are of little interest by themselves, but are introduced here to aid in comparing local regression with other methods. In practice, one should think directly in terms of the basis of local polynomials being fitted.

Figure 4 illustrates the local regression method. A straight line is fitted to the data on the window $[0.5, 0.7]$; since this line closely approximates the true mean, the bias in estimating $f(0.6)$ is small. Moreover, the effective kernel has a smooth form, and hence the local regression estimate is less variable than the GM estimate. The bias reduction of the local regression method is particularly advantageous in boundary regions; compare Figures 5 and 3.

Figures 4,5 about here.

When slope effects are properly modeled, the main source of bias is curvature of the mean function. If local quadratic (or higher order) polynomials are fitted, further reduction in bias is obtained. However, fitting higher order polynomials generally gives a more variable estimate. For practical applications, local linear and local quadratic fitting are usually the most useful procedures.

Stone (1980, 1982) studies rates of convergence in nonparametric regression and shows local regression achieves rates that are optimal in a certain minimax sense. Müller (1987) establishes an asymptotic equivalence with kernel methods in a very restrictive setting. Cleveland and Devlin (1988) study local regression for multivariate predictors. Fan (1992a,b) studies local linear regression and establishes some asymptotic minimax efficiency properties.

Local polynomials are a popular choice amongst data analysts. Cleveland's *lowess* implementation in the language S (Becker, Chambers and Wilks, 1984) is widely used; it features a hybrid near-neighbor kernel and allows down-weighting of outliers. A multidimensional version is available in the 1991 release of S, and is described in Cleveland, Grosse and Shyu (1991).

Despite the intuitive appeal and excellent asymptotic properties of local regression, the methods are not always well understood. In recent years, methods involving special classes of kernels, and modifications to the basic kernel methods, have been popular bias reduction techniques.

The purpose of this paper is to argue that for routine use, the local regression technique offers many advantages over modified kernel methods. Figures 4 and 5 provide a simple illustration as to how local polynomials model slope effects; curvature effects are modeled similarly if higher order polynomials are fitted. By contrast, there is no analog of these figures to explain how methods such as high order kernels and boundary kernels work.

A second advantage is relative insensitivity to the design. While some assumptions, such as a smooth design density, will usually be required in asymptotic analysis of nonparametric regression estimates, such assumptions are not always appropriate for designs encountered in practice. A nonparametric regression estimate should continue to perform well for unusual designs. Methods that perform poorly for unusual designs or are inefficient for random designs are unsuitable for general purpose use, such as in statistical software. The local linear regression models slope effects under just one design assumption: that the inverse in (3) exists.

Closely related to the problem of modeling slopes and curvature is the problem of derivative estimation. As a byproduct of the local linear fitting, we obtain derivative estimates: namely the slope of the local line. If higher order polynomials are fitted, estimates of higher order derivatives are also obtained.

Perhaps the biggest advantage of local regression is when the predictor is two or three dimensional. In this case, a kernel estimate may be influenced by boundary effects over much of the domain, and much structure may be lost by ignoring the effects. Adaptation of the local regression estimate to multivariate predictors is simple: we just change the basis functions in (3).

The performance of regression estimates is often characterized by mean squared error and other measures of accuracy. We do not claim that local regression methods will always have dramatic advantages over modified kernels under such measures. Indeed, in some circumstances modified kernel methods closely approximate local regression methods and mean squared error will be very similar. Rather, we argue that local regression provides a simple, intuitive and automatic solution to the problems modified kernel methods are attempting to address.

The remainder of this paper contains a more detailed discussion of how local polynomial methods provide a solution to various problems, and provides comparisons with kernel and modified kernel methods. The problem of unequally spaced observations is discussed in Section 2; boundary effects are discussed in Section 3 and derivative estimation in Section 4. The multivariate case is discussed in Section 5.

The polynomial smoothing spline is another popular smoothing technique; see Silverman (1985). It is the solution to an optimization problem and adapts to many of the bias problems associated with kernel estimates. While kernel smoothers permit reasonably straightforward asymptotic analysis, splines seem to simplify the generation and analysis of algorithms (Buja, Hastie and Tibshirani, 1989; Hastie and Tibshirani, 1990). The smoothing spline and local regression methods appeal to rather different intuitive interpretations, and it is unlikely that either will have universally dominant performance.

2 Unequally Spaced Observations

Using the linearity of kernel and local regression methods, and using a series expansion of f around x, we obtain

$$\begin{aligned} E\hat{f}(x) &= \sum_{j=1}^{n} l_j(x) f(x_j) \\ &= f(x)\sum_{j=1}^{n} l_j(x) + f'(x)\sum_{j=1}^{n}(x_j - x) l_j(x) \\ &\quad + \frac{f''(x)}{2}\sum_{j=1}^{n}(x_j - x)^2 l_j(x) + R. \end{aligned} \tag{5}$$

The bias is defined as $E\hat{f}(x) - f(x)$. Under some regularity assumptions, the locality of the kernel implies the remainder term R is small for both kernel and local linear regression estimates. It is therefore of interest to examine the terms involving $f'(x)$ and $f''(x)$ in more detail.

Although (5) is for a fixed design, the expansion continues to hold for a random design if we first condition on the observed x_i. A similar expression for the variance is

$$\operatorname{var}\hat{f}(x) = \sum_{i=1}^{n} l_j(x)^2 \operatorname{var}\epsilon_j = \sigma^2 \|l(x)\|^2 \tag{6}$$

if the residuals have constant variance σ^2.

For the Nadaraya-Watson estimate (1), the coefficient of $f'(x)$ in (5) is

$$\sum_{j=1}^{n}(x_j - x) l_j(x) = \frac{\sum_{j=1}^{n}(x_j - x)K(x - x_j)}{\sum_{j=1}^{n} K(x - x_j)}. \tag{7}$$

If the kernel is symmetric and observations are symmetrically distributed around x, this equals 0. However, if the observations are asymmetric, this term will in general be nonzero, and slope of the mean function causes a biased estimate.

Alternatively, consider the local regression estimate (3) with basis functions $\{1, x, \ldots, x^q\}$ for some $q \geq 1$. If $p(x)$ is in the linear span of the basis functions (i.e. p is a polynomial of degree $\leq q$), then it follows from the definition (3) that

$$p(x) = \sum_{j=1}^{n} l_j(x) p(x_j) \tag{8}$$

for all x. We are fitting a polynomial regression of degree q (by weighted least squares) to a set of points lying exactly on a polynomial of degree q or lower. As long as there are at least q nonzero weights, the fit will be exact, and (3) gives the fitted value at x, namely $p(x)$.

It follows from (5) that all bias terms of degree q or lower are automatically removed, regardless of the design. To see this, let $p(x) = (x - z)^k$ in (8), and thus $0 = p(z) = \sum_{j=1}^{n} l_j(z)(x_j - z)^k$ for all $k \leq q$. In particular, when a local linear regression ($q = 1$) is used, the dependence of bias on the slope of f is removed for all x.

An alternative method of adjusting for the effect of unequally spaced observations is to consider proximity to other observations when assigning weights. Relatively low weight is given to observations occurring in clumps. The original method of this type was proposed by Priestley and Chao (1972), while the GM estimate (2) is presently popular. If the kernel is compactly supported on $[-h, h]$, the term in (5) involving $f'(x)$ is approximately eliminated in the interior region:

$$\sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} (x_i - x) K(x - u) du \approx \int_{x-h}^{x+h} (u - x) K(x - u) du = 0.$$

The GM and local linear regression methods have been illustrated in Figures 2 and 4; both are seen to have similar bias reduction. It is also of interest to compare the variance of the estimates. Using (6), $\mathrm{var} \hat{f}(0.6) = 0.059\sigma^2$ for the NW estimate; $0.070\sigma^2$ for the local linear estimate and $0.083\sigma^2$ for the GM estimate. In both cases the bias reduction is accompanied by an increase in variance; as expected, the GM estimate is more variable than the local regression.

Further insight is gained from the asymptotic analysis carried out by other authors. Assuming the design density is continuous and bounded away from 0, Fan (1992a) shows the local linear regression has the same asymptotic variance as the NW estimate: the price paid for the bias reduction is minimal.

When f is assumed to be in a class of twice differentiable functions, Fan also derives some asymptotic minimax properties for the local regression.

Jennen-Steinmetz and Gasser (1988) and Gasser and Engel (1990) show the GM estimate has an asymptotic variance 1.5 times as large as the NW estimate. When the observations are unequally spaced, the GM estimate is an inefficient way to model slope effects.

Chu and Marron (1991) suggest improving the variance properties of the GM estimate by alternative methods of specifying s_i. With appropriate choices of s_i it is quite probable that one can coerce the effective kernel in Figure 2 to a form similar to Figure 4; however, the effort seems unnecessary since the local regression has automatically achieved the desired result.

When the linear terms are properly modeled, the bias expansion (5) will be dominated by curvature terms. If local quadratic ($q = 2$) fitting is used, then dependence of the bias on f'' is removed.

In some cases curvature effects can be approximately modeled using special classes of higher order kernels; see Gasser and Müller (1979). This approach is less intuitive than directly modeling curvature using local quadratic regression, and continues to be inefficient for random designs. Use of higher order kernels with the NW estimate can result in instability, similar to that illustrated with boundary kernels in the next section.

Of course, one can fit higher order polynomials or kernels and obtain further bias reduction. The downside is that higher order fits are more variable; the selection of order can be addressed as a bias-variance trade-off.

The message of this section is that local regression provides a simple and intuitive way to correct biases to any given order, and performs as well as or better than other methods of a comparable order.

3 Boundary Effects

At a boundary point, Figure 3 shows slope of the mean function induces particularly severe bias in the Nadaraya-Watson estimate. This can also be seen from (7). Since all the $x_j - x$ have the same sign, there is no cancellation of terms in the numerator when a positive kernel is used.

Local linear regression is shown in Figure 5; this provides a simple and intuitive way of modeling slopes in the boundary region. However, the local regression will be substantially more variable, and unlike the situation discussed in the previous section, the variance increase persists even asymptotically.

Do we gain by modeling slopes in the boundary region? Suppose the observations are uniformly distributed on $[0,1]$. With appropriate bandwidths, an asymptotic analysis similar to that in Stone (1982) shows $\hat{f}(0)$ has mean squared error $O(n^{-2/3})$ for the NW estimate and $O(n^{-4/5})$ for the local linear regression. This suggests that for large n, fitting slopes is beneficial.

For small n, the situation is more difficult; either estimate may have better MSE. Suppose our purpose in fitting a nonparametric regression curve is to uncover structure in the true mean. Generally, structure that can be found by the NW estimate but can't be found by the local linear regression will be confounded with boundary bias, and it would be unusual for the NW estimate to identify structure that the local regression can't detect. However, if prediction is our main interest, then bias and variance are the major considerations, and the NW estimate may give better predictions in situations without much structure.

Several methods have been proposed to modify kernel estimators to handle boundary effects. A popular method is through the use of special boundary kernels. If the data are uniformly distributed on $[0,1]$, (7) suggests imposing the constraints

$$\begin{aligned} \int_0^1 K_x(x-u)du &= 1 \\ \int_0^1 (u-x)K_x(x-u)du &= 0 \end{aligned} \tag{9}$$

Typically, the modified kernels are of the form

$$K_x(x-u) = K(x-u)(\gamma_0 + \gamma_1(x-u)) \tag{10}$$

where γ_0 and γ_1, dependent on x, are determined by the moment conditions (9); see Müller (1991).

The effective kernels for the local linear regression also have the form (10), if the integrals in (9) are replaced by sums over the observed x_j. This implies that boundary kernels are essentially an attempt to directly approximate the effective kernels from local linear regression; under appropriate conditions there is an asymptotic equivalence between the two approaches. This is investigated in more detail by Müller (1988, section 4.3), and references therein.

What problems exist with the use of boundary modified kernels? First, there is no intuitive explanation as to why they work; unlike the local regression method which has been simply illustrated in Figure 5.

A more serious problem is a lack of design adaptability. Suppose the design density is $2xI_{[0,1]}(x)$. If boundary kernels satisfying (9) are used to estimate $f(0)$, the denominator of (1) is close to 0, and anything could result.

Clearly, we could replace (9) by conditions appropriate to this type of density, and derive more classes of boundary kernels. However, either user or software is then required to decide what type of boundary kernel is appropriate, according to an observed density of points near the boundary. By contrast, the use of local regression automatically adjusts to the various types of density.

Boundary kernels can also be combined with integral kernels; see for example Müller (1991). This reduces the problem of design adaptability associated with the NW estimate. However, our objection to integral kernel methods is clear from previous sections: they are inefficient for random designs.

Other methods to reduce boundary bias have been proposed, including extrapolation methods (Rice, 1984) and reflection methods (Hall and Wehrly, 1991).

Extrapolation methods involve combining two different kernel estimates with different bandwidths, to eliminate the $f'(x)$ terms. Suppose the estimates are $\hat{f}_l(x)$ and $\hat{f}_m(x)$ with

$$E\hat{f}_l(x) = f(x) + f'(x)L + R_l; \qquad E\hat{f}_m(x) = f(x) + f'(x)M + R_m.$$

Here, L and M depend on x and the design points, but not f. An appropriate linear combination of the two estimates removes the $f'(x)$ terms. The difficulty is that we need to choose several smoothing parameters; the number increasing depending on how many terms we wish to remove. It is also difficult to see exactly how this method operates on the data.

The reflection technique proposed by Hall and Wehrly (1991) involves generating pseudo-data, so that in an enlarged data set the boundaries of the original data set are now on the interior. If the x_i are supported on $[a,b]$, the single point $f(a)$ is estimated using a one-sided boundary kernel, and additional data generated by reflecting in this estimate. Similar reflection is carried out at b. The estimate $\hat{f}(x)$ for $x \in [a,b]$ is then constructed using an ordinary kernel estimate. However, we have shown above that boundary kernels can fail for some designs, and hence this reflection technique can also fail.

There are two main conclusions of this section. Firstly, the NW estimate may have substantial boundary bias, and it is usually preferable to use methods which model slopes in boundary regions. Secondly, the local linear regression is the best method by which to model slope effects. The method is automatic, has a simple intuitive interpretation and adapts well to different designs.

4 Derivative Estimation

The first and second derivatives of a regression function often have important physical interpretations and it is interesting to study estimates of these quantities. A linear estimate of $f''(x)$ is

$$\hat{f}''(x) = \sum_{j=1}^{n} d_j(x) Y_j$$

for appropriate d_j. In view of the bias expansion (5), a minimal set of constraints is

$$\begin{aligned} \sum_{j=1}^{n} d_j(x) &= 0 \\ \sum_{j=1}^{n} (x_j - x) d_j(x) &= 0 \\ \sum_{j=1}^{n} (x_j - x)^2 d_j(x) &= 2. \end{aligned} \tag{11}$$

Suppose we fit a local polynomial at x of degree $q \geq 2$, and take $\hat{f}''(x)$ to be the second derivative of the local polynomial (*not* the second derivative of $\hat{f}(x)$). Then

$$d(x)^T = b(x)''^T (B^T W(x) B)^{-1} B^T W(x) \tag{12}$$

and $d(x)^T B = b(x)''^T$. This implies $\sum_{i=1}^{n} d_j(x) p(x_j) = p''(x)$ for any polynomial of degree $\leq q$, and hence the conditions (11) are satisfied. One can go further and play the bias reduction game by fitting even higher order polynomials. Rates of convergence for this type of derivative estimate are studied by Stone (1980).

Gasser and Müller (1984) propose integral kernel type derivative estimates,

$$\hat{f}''_{GM}(x) = \sum_{j=1}^{n} Y_j \int_{s_{j-1}}^{s_j} K^{(2)}(x-u)du$$

where $K^{(2)}$ satisfies

$$\begin{aligned} \int K^{(2)}(u)du &= 0 \\ \int uK^{(2)}(u)du &= 0 \\ \int u^2 K^{(2)}(u)du &= 2. \end{aligned}$$

Under suitable assumptions, (11) will be approximately satisfied by these weights, and for well behaved designs the results of Müller (1987) establish an asymptotic equivalence with the local regression method. However, under random designs $\hat{f}''_{GM}$ will again be more variable.

5 Multivariate Smoothing

There are many important applications of smoothing techniques in two or more dimensions. Examples include images from medical scanning devices and satellite photographs, as well as geographically recorded data. Smoothing becomes less feasible as the dimension of the predictors x_i gets too large, due to *curse of dimensionality* (Bellman, 1961). Also, beyond two or three dimensions, a full nonparametric regression surface is difficult to visualize, and regression methods which try to capture lower dimensional structure may be appropriate. Methods of this type include projection pursuit (Friedman and Stuetzle, 1981) and additive models (Hastie and Tibshirani, 1990).

The definition of local regression is easily extended to multiple predictors and the methods have been successfully applied in two and three dimensions; see Cleveland and Devlin (1988) for examples.

Many properties of local regression estimates extend immediately from the one dimensional case. For example, a local linear regression will model the slope of the mean function; the bias depends only on the second and higher order partial derivatives of the true mean function. Asymptotic properties such as rates of convergence can be derived in a straightforward manner; see for example Stone (1980).

The local regression method is particularly useful with unusual designs. For example, Buta (1987) made velocity measurements on a galaxy at positions on the celestial sphere. Because of the way measurements were made, the predictor variables form a star-shaped pattern; see Figure 6. Cleveland and Devlin (1988) fitted a local quadratic model to this dataset, and found revealing structure in the velocity measurements.

Figure 6 about here.

Could modified kernel methods be devised to adequately model multivariate regression functions? We would need modifications to handle boundary effects, and nonuniformity and curvature effects in the interior region.

If a boundary region can be precisely specified, then boundary kernels can be derived using multivariate extensions of (9) and (10). However, the precise specification of a boundary region is difficult; for the design in Figure 6 it certainly is not clear where the boundary should be. Also, the boundary region may be quite complex making the evaluation of the extensions of (9) complicated. Finally, the success of boundary kernels requires that predictors be approximately uniformly distributed near boundary points.

For nonuniform observations, extension of Gasser-Müller type estimates to the multivariate setting is discussed by Ahmad and Lin (1984) and Müller (1988). This involves dividing the predictor space into sets $A_{i,n}, i = 1, \ldots, n$ associated with each observations and using the natural extension of the Gasser-Müller estimate. However, unless the design is a grid of points, deriving the sets $A_{i,n}$ is fairly arbitrary and integrating the kernel over these sets is an additional complication. As in one dimension, integral kernel estimates will have poor variance properties.

We have indicated the severe difficulties encountered when trying to modify kernel methods for multivariate designs. Supposing these can be overcome, it seems the best we can expect is modified kernels which closely approximate the effective kernels of local regression methods. Clearly, the local regression method is preferable, since the bias is corrected automatically and simultaneously for:

- asymmetric neighborhoods in the interior;
- closeness to the boundary, and
- the shape of the boundary.

6 Examples

This section contains a numerical study of some of the methods discussed. We restrict ourselves to second order methods; that is, methods that model the slope of the mean function.

Three different designs are considered:

1) 50 points equally spaced on $[0,1]$, $x_i = (i-1)/49$.

2) 50 points on $[0,1]$, $x_i = (1-\sqrt{1-2(i-1)/49})/2$ if $i \leq 25$ and $x_i = (1+\sqrt{2(i-1)/49-1})/2$ if $i \geq 26$.

3) Random design, 50 points from density $6x(1-x)I_{[0,1]}(x)$.

The mean function considered is $f(x) = x^2$.

We use the Epanechnikov kernel $K(x) = (1-x^2)I_{[-1,1]}(x)$, and the boundary modified version given in Table 1 of Müller (1991). A varying bandwidth is used:

$$h = h(x) = \begin{cases} 0.6 - x & x < 0.3 \\ 0.3 & 0.3 \leq x < 0.7 \\ x - 0.4 & 0.7 \leq x \end{cases}$$

which has reasonable sized boundary and central regions for illustration. When the n data points are equally spaced on $[0,1]$, this corresponds roughly to a $0.6n$ nearest neighbor bandwidth.

We consider bias and variance separately. For convenience, the residual variance is taken to be 1. In tables 1 and 2, the five estimates are summarized by their integrated variances and squared biases. For three of the estimates, bias and variance are compared on a pointwise basis in Figure 7.

Tables 1,2; Figure 7 about here.

In the central region, all estimates had comparable performance for the uniform design, as expected. The second design has a low density of points in the central region, and the NW estimate has substantially larger bias than the other methods, but slightly less variable. For the third design, the non-uniformity again results in the NW estimate having slightly larger bias. Since this design is random, the GM estimate is more variable than the local regression.

The boundary regions are more interesting. As expected, the local linear regression and the two boundary modified estimates are more variable and less biased than the unmodified estimates. Since $f'(0) = 0$, most of the bias reduction occurs at the upper boundary. The modified NW estimate achieves little bias reduction for design 2, and is unstable for design 3. The boundary modified GM estimate has greater bias and greater variance than the local linear regression, even for the uniform design where we would expect the methods to be similar. However, the 'optimum' boundary kernels given by Müller (1991) and applied here give low weights to observations near the boundary.

In summary, the local linear regression has performed well for all three designs. Although more variable than the Nadaraya-Watson estimate in the boundary regions, there is a substantial bias reduction, particularly at the upper boundary. The Gasser-Müller estimates performs comparably to the local linear regression in terms of bias correction over the central region, although is more variable. In the boundary regions, the local linear regression outperforms the other boundary corrected methods.

7 Conclusions and Discussion

Kernel smoothing, especially the Nadaraya-Watson estimate (1), has great intuitive appeal, and is easily motivated. There are however bias problems, especially at boundaries. In practice boundary effects may occur over a substantial region, especially in dimensions greater than one, and the bias of the NW estimate may mask interesting structure. We have shown that, unlike the modified kernel approaches, local polynomial smoothing attends to the bias problems while retaining the original simplicity. Local polynomials generalize immediately to smoothing problems in higher dimensions, and their bias correction properties accompany them.

Another form of bias commonly encountered in practice curvature esffects, often referred to as "trimming the hills and filling the valleys". This is particularly noticeable when the signal-to-noise ratio is very high. In this case local-quadratic smoothers perform well; once again an attractive alternative to higher order kernels.

Kernel smoothing can be extended to nonparametric regression in likelihood based models ("Local Likelihood", Hastie and Tibshirani, 1990, especially sections 6.5.1 and 6.13 and references therein); once again bias problems can occur, and can be corrected by fitting polynomials locally.

ACKNOWLEDGEMENTS

We thank Bill Cleveland, Colin Mallows and Daryl Pregibon for comments on an earlier draft, and Jianqing Fan for supplying us with preprints of his work.

REFERENCES

Ahmad, I. A. and Lin, P.-E. (1984). Fitting a multiple regression function. *J. Statist. Plan. Inf.* **9**, 163-176.

Becker, R., Chambers, J. and Wilks, A. (1984). *The New S Language.* Wadsworth, Pacific Grove, California.

Bellman, R.E. (1961). *Adaptive Control Systems.* Princeton University Press, Princeton.

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17**, 453-555.

Buta, R. (1987). The structure and dynamics of ringed galaxies. III. Surface photometry and kinematics of the ringed nonbarred spiral NGC7531. *Astrophysical J. Supp. Ser.* **64**, 1-37.

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

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assn.* **74**, 829-836.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *J. Amer. Statist. Assn.* **83**, 596-610.

Cleveland, W. S; Grosse, E. and Shyu, W. M. (1991). Local Regression Models. *Statistical Models in S.* Ed. Chambers and Hastie. Wadsworth,

Pacific Grove, California.
FAN, J. (1992a). Design-adaptive nonparametric regression. *J. Amer. Statist. Assn.* **87**, 998-1004.
FAN, J. (1992b). Local linear regression smoothers and their minimax efficiency. *Ann. Statist.*
FRIEDMAN, J. and STUETZLE, W. (1981). Projection Pursuit Regression. *J. Amer. Statist. Assn.* **76**, 817-823.
GASSER, TH. and ENGEL, J. (1990). The choice of weights in kernel regression estimation. *Biometrika* **77**, 377-381.
GASSER, TH. and MÜLLER, H.-G. (1979). Kernel estimation of regression functions. *Lecture Notes in Mathematics* **757**, 23-68. Springer-Verlag, Berlin.
GASSER, TH. and MÜLLER, H. G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.* **11**, 171-185.
HALL, P. and WEHRLY, T. E. (1991). A geometrical method for removing edge effects from kernel-type nonparametric regression estimates. *J. Amer. Statist. Assn.* **86**, 665-672.
HASTIE, T. and TIBSHIRANI, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.
JENNEN-STEINMETZ, C. and GASSER, TH. (1988). A unifying approach to nonparametric regression estimation. *J. Amer. Statist. Assn.* **83**, 1084-1089.
MÜLLER, H.-G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *J. Amer. Statist. Assn.* **82**, 231-238.
MÜLLER, H.-G. (1988). Nonparametric regression analysis of longitudinal data. *Lecture Notes in Statistics* **46**, Springer Verlag, Berlin.
MÜLLER, H.-G. (1991). Smooth optimum kernel estimators near endpoints. *Biometrika* **78**, 521-530.
NADARAYA, E. A. (1964). On estimating regression. *Theor. Probab. Appl.* **9**, 141-142.
PRIESTLEY, M. B. and CHAO, M. T. (1972). Nonparametric function fitting. *J. R. Statist. Soc.* Ser. B **34**, 384-392.
RICE, J. (1984). Boundary modification for kernel regression. *Comm. Statist. Theor. Meth.* **13**, 893-900.
SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *J. R. Statist. Soc.* Ser. B **47**, 1-52.
STONE, C. J. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.* **5**, 595-620.

STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8**, 1348-1360.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1040-1053.

WATSON, G. S. (1964). Smooth regression analysis. *Sankhya* Ser. A , **26**, 359-372.

Estimator	Design 1 Bound.	Design 1 Center	Design 2 Bound.	Design 2 Center	Design 3 Bound.	Design 3 Center
NW	0.02310	0.01633	0.02429	0.02615	0.02232	0.01309
GM	0.01346	0.01631	0.01924	0.04676	0.02722	0.01934
Loc Lin	0.04127	0.01633	0.03584	0.03363	0.07631	0.01372
Mod NW	0.05453	0.01633	0.03097	0.02616	47.6342	0.01309
Mod GM	0.05407	0.01631	0.06006	0.04676	0.10885	0.01934

Table 1: Integrals of $\|l(x)\|^2$ over the boundary region $[0, 0.3] \cup [0.7, 1.0]$ and the center region $[0.3, 0.7]$.

Estimator	Design 1 Bound.	Design 1 Center	Design 2 Bound.	Design 2 Center	Design 3 Bound.	Design 3 Center
NW	0.01126	0.00013	0.00513	0.00267	0.01919	0.00026
GM	0.04314	0.00013	0.04297	0.00015	0.04450	0.00013
Loc Lin	0.00020	0.00013	0.00009	0.00020	0.00056	0.00011
Mod NW	0.00055	0.00013	0.00426	0.00266	8.88807	0.00026
Mod GM	0.00054	0.00013	0.00057	0.00015	0.00064	0.00014

Table 2: Integrated squared biases over the boundary region $[0, 0.3] \cup [0.7, 1.0]$ and the center region $[0.3, 0.7]$.

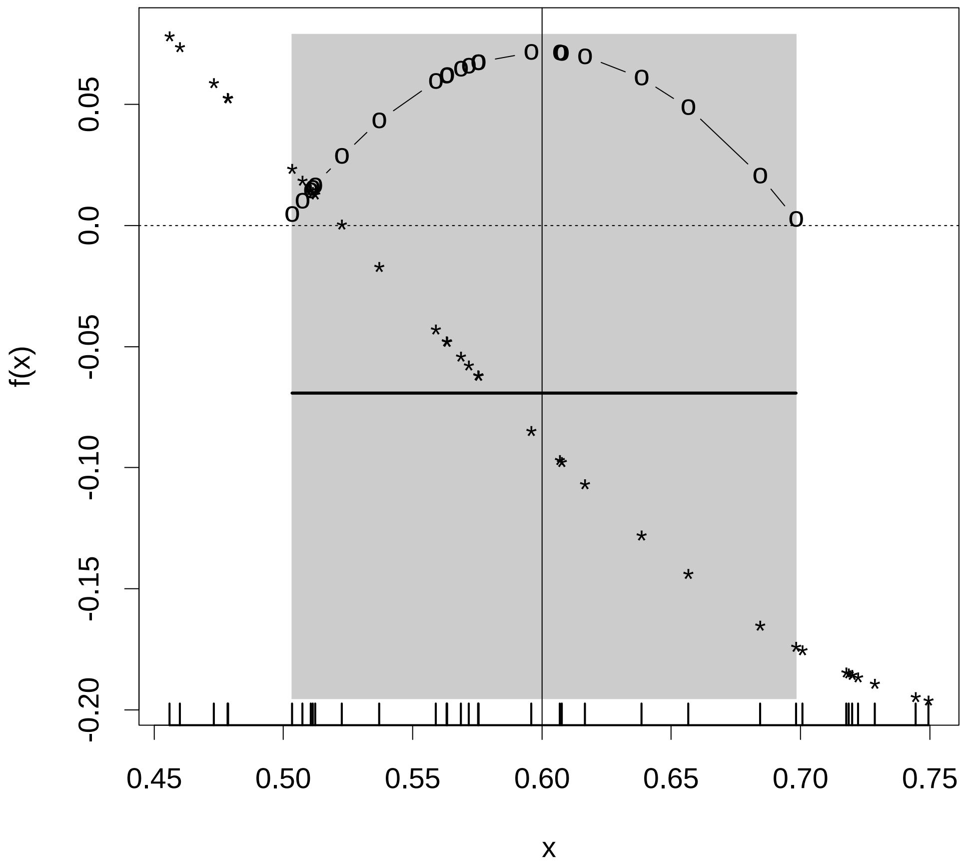

Figure 1: *Effect of Asymmetry on the Nadaraya-Watson estimator. Suppose we observe the data indicated by the asterisks; for clarity shown with no noise. We estimate $f(0.6)$ using the locally-constant NW fit (thick line) using Epanechnikov's kernel $K(x/10) = (1 - x^2)I_{[-1,1]}(x)$, indicated by the circles. The asymmetry of observations causes substantial bias.*

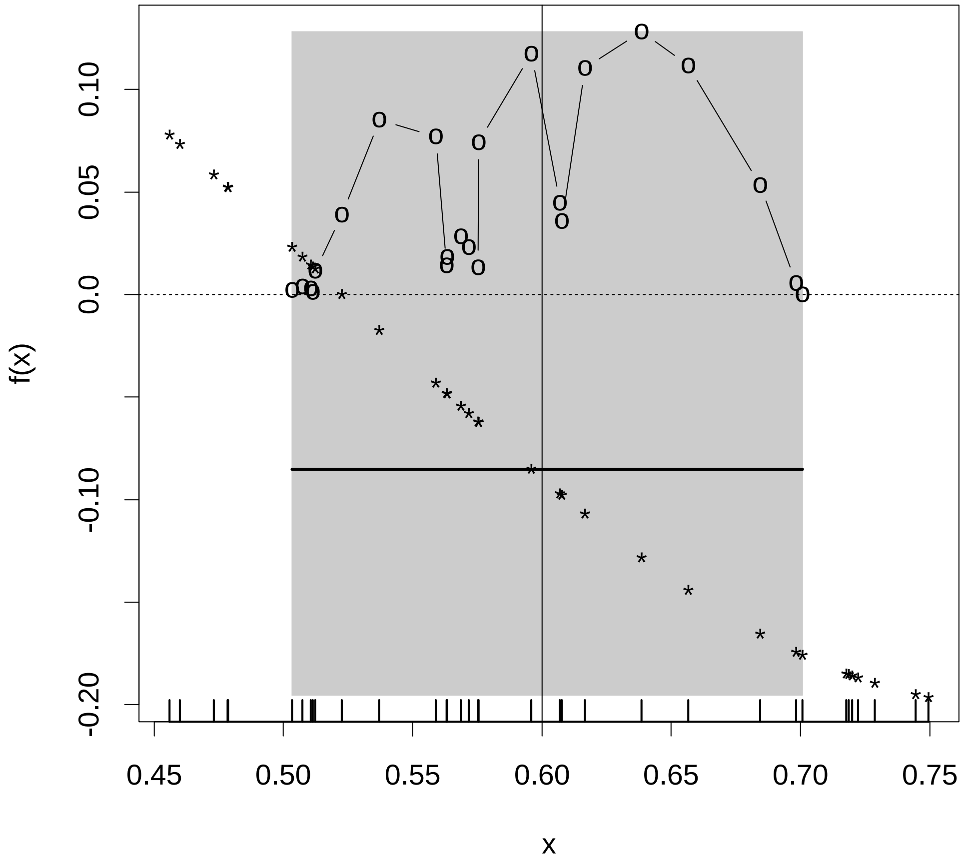

Figure 2: *Gasser-Müller estimate. The use of integral kernels downweights some of the clustered observations in* $[0.5, 0.6]$*, substantially improving the bias. However, the noise of the effective weights introduces extra variability when the data is observed with error.*

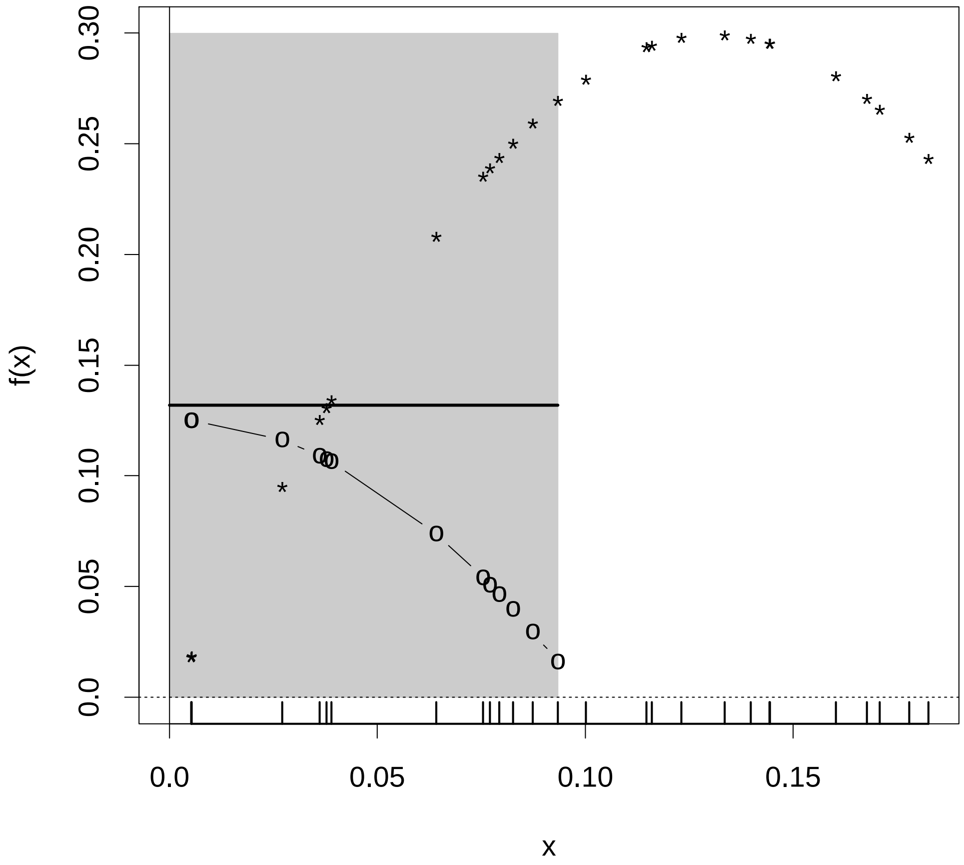

Figure 3: *Nadaraya-Watson estimate, boundary effects. When the x_i are in the interval $[0, 1]$ and we attempt to estimate $f(0)$, slope of the mean function induces particularly severe bias.*

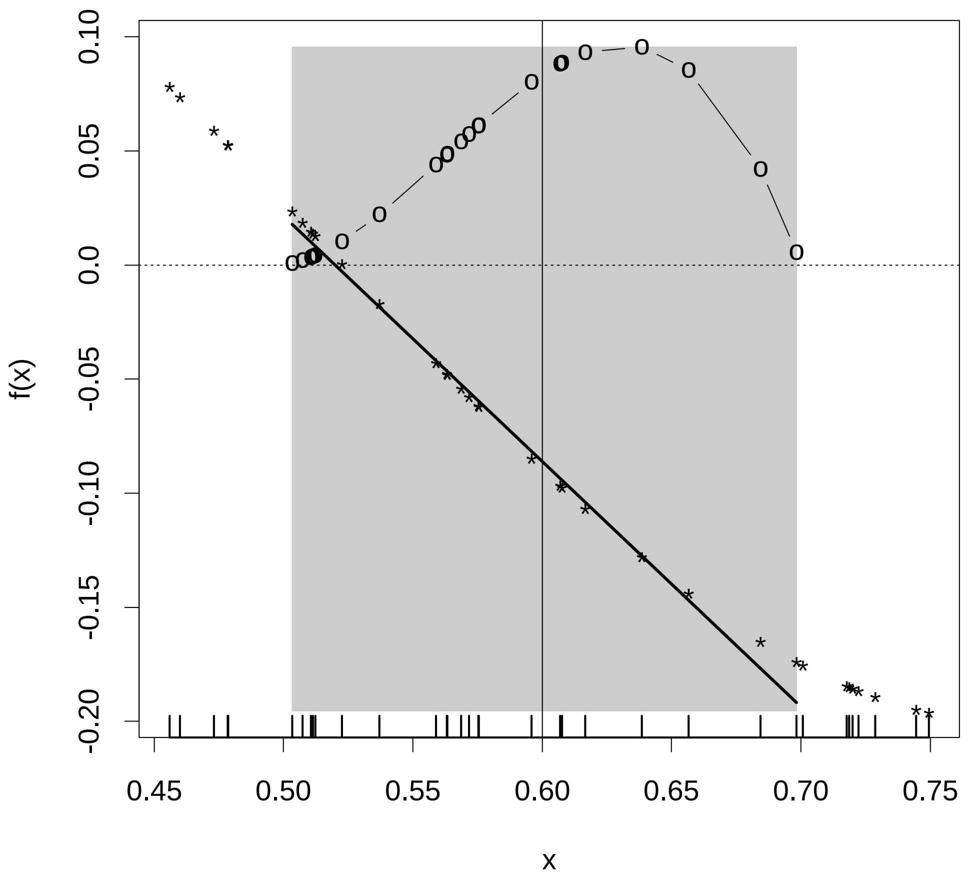

Figure 4: *Local linear regression. We estimate $f(0.6)$ using weighted least squares with weights assigned by Epanechnikov's kernel. This has substantially reduced the bias associated with the NW estimate. Moreover, the effective weights, shown by the circles, do not have the noisy behavior associated with the GM estimate.*

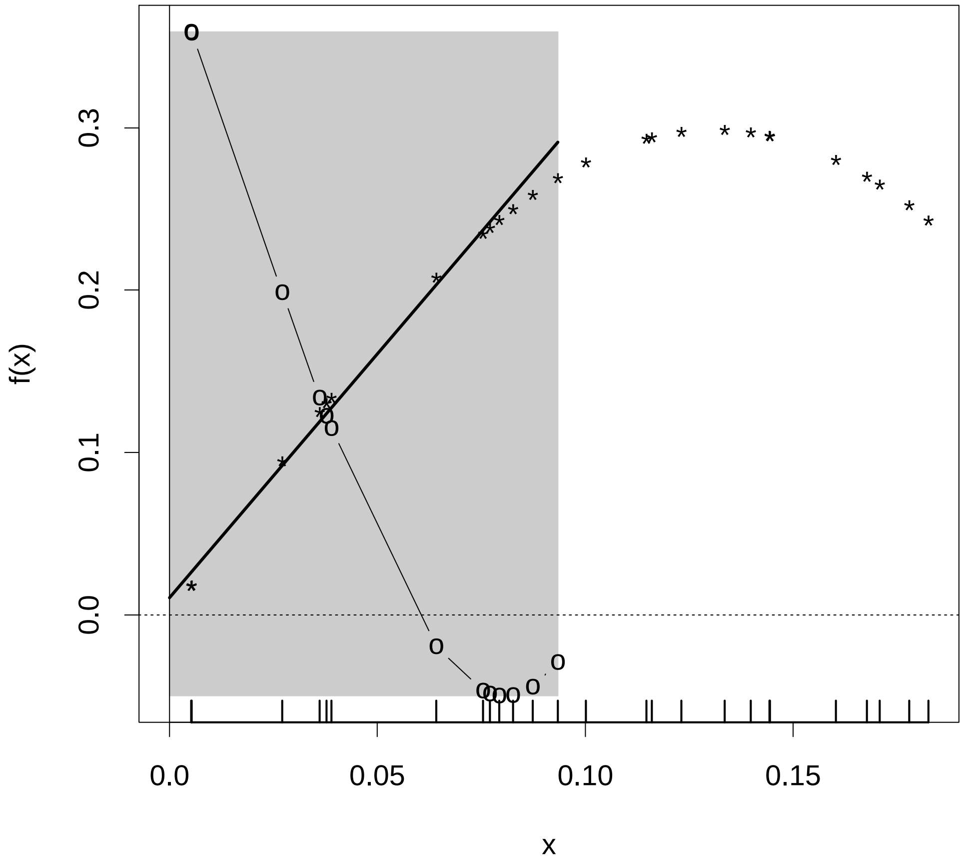

Figure 5: *Local linear regression, boundary region. Fitting the weighted least squares has substantially reduced the bias of the Nadaraya-Watson estimate in the boundary region.*

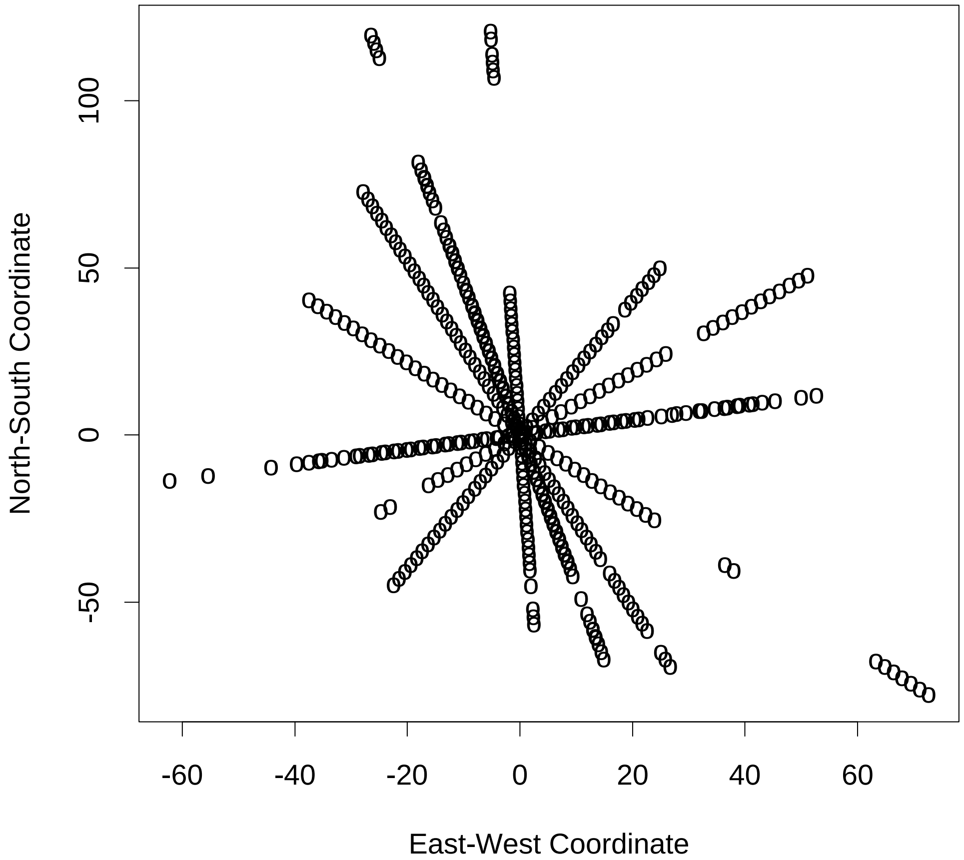

Figure 6: *Predictor variables for the NGC7531 dataset (Buta, 1987; Cleveland and Devlin, 1988).*

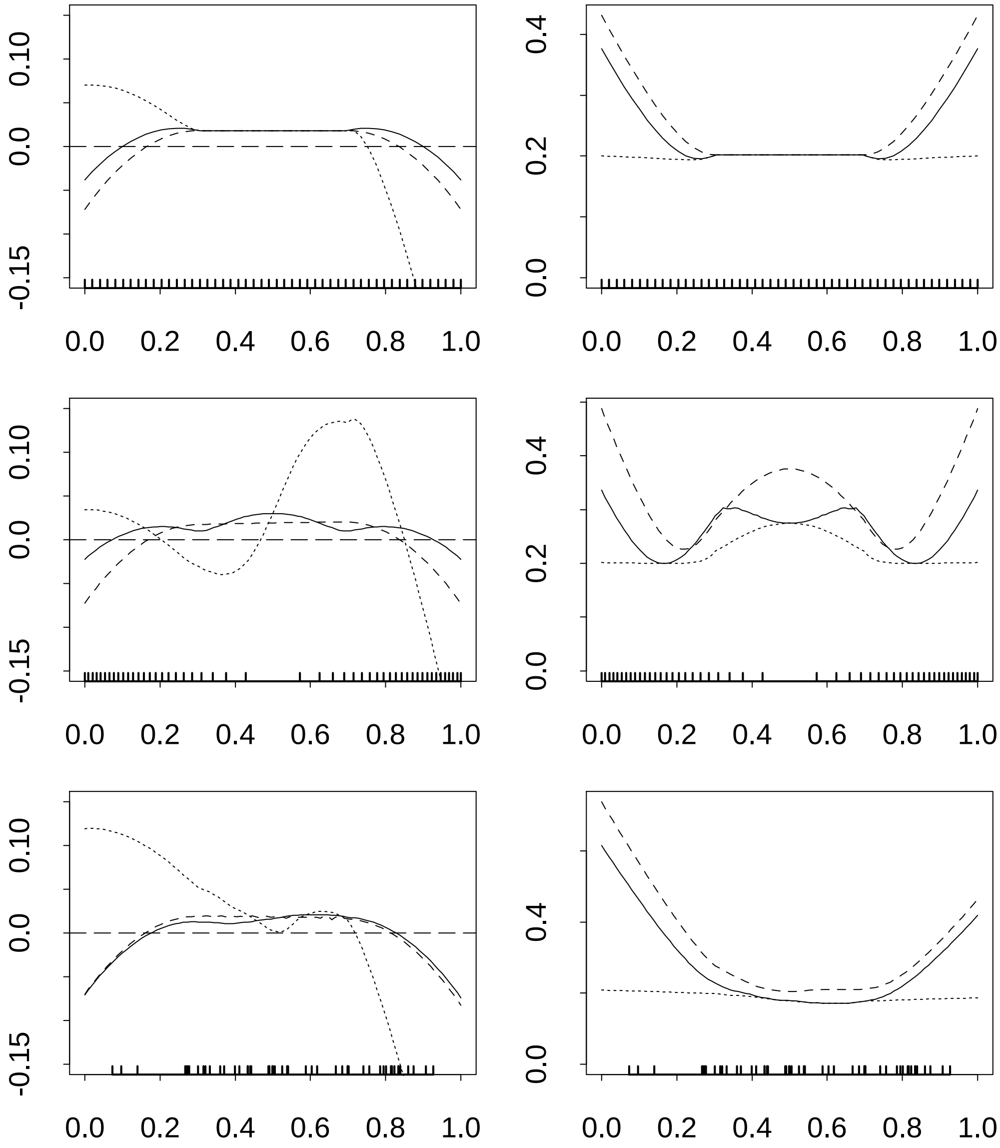

Figure 7: *Comparisons of biases (left column) and* $\|l(x)\|$ *(right column) for three designs. The solid lines represent a local linear regression; the short dashes represent a Nadaraya-Watson estimate and medium dashes are the boundary corrected Gasser-Müller estimate. The long dashes show zero bias for reference.*