Design-adaptive Nonparametric Regression

JIANQING FAN*

In this article we study the method of nonparametric regression based on a weighted local linear regression. This method has advantages over other popular kernel methods. Moreover, such a regression procedure has the ability of design adaptation: It adapts to both random and fixed designs, to both highly clustered and nearly uniform designs, and even to both interior and boundary points. It is shown that the local linear regression smoothers have high asymptotic efficiency (i.e., can be 100% with a suitable choice of kernel and bandwidth) among all possible linear smoothers, including those produced by kernel, orthogonal series, and spline methods. The finite sample property of the local linear regression smoother is illustrated via simulation studies. Nonparametric regression is frequently used to explore the association between covariates and responses. There are many versions of kernel regression smoothers. Some estimators are not good for random designs, such as in observational studies, and others are not good for nonequispaced designs. Furthermore, most nonparametric regression smoothers have "boundary effects" and require modifications at boundary points. However, the local linear regression smoothers do not share these disadvantages. They adapt to almost all regression settings and do not require any modifications even at boundary. Besides, this method has higher efficiency than other traditional nonparametric regression methods.

KEY WORDS: Boundary effects; Kernel estimator; Linear smoother; Local linear regression; Minimax efficiency.

1. INTRODUCTION

Consider bivariate data that can be thought of as a random sample from a certain population. It is common practice to study the association between covariates and responses via regression analysis. Nonparametric regression provides a useful explanatory and diagnostic tool for this purpose. See Eubank (1988), Härdle (1990), and Müller (1988) for many examples of this and good introductions to the general subject area.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a population having a density $f(x, y)$. Let $f_X(x)$ be the marginal density of X. Denote the regression function by $m(x) = E(Y|X = x)$ and the conditional variance function by $\sigma^2(x) = \text{var}(Y|X = x)$. Several methods have been proposed for estimating $m(x)$: kernel, spline, and orthogonal series methods. Among these are two popular kernel methods proposed by Gasser and Müller (1979), Nadaraya (1964), and Watson (1964). With K being a kernel and h_n being a bandwidth, Table 1 summarizes the asymptotic behavior of the Nadaraya–Watson estimator (3.4), the Gasser–Müller estimator (3.5), and the local linear (regression) smoother (2.2) to be introduced in Section 2.

The bias of the Nadaraya–Watson estimator depends on the intrinsic part m'' interplaying with the artifact $m'f'_X/f_X$ due to the local constant fit. Keeping $m''(x)$ fixed, we first remark that in the highly clustered design where $|f'_X(x)/f_X(x)|$ is large, the bias of the Nadaraya–Watson estimator is large. Thus this estimator cannot adapt to highly clustered designs. Note also that when $|m'(x)|$ is large, so is the bias of that estimator. Thus, even in the situation of linear regression $m(x) = a + bx$ with a large coefficient b, the bias of the estimator is also large. In other words the Nadaraya–Watson estimator is not good at testing linearity. See Chu and Marron (1991) for further discussion.

For random designs, the Gasser–Müller estimator has an asymptotic variance 1.5 times as large as that of the Nadaraya–Watson estimator (Chu and Marron 1991; Mack and Müller 1988). On the other hand the bias of the Gasser–Müller estimator is simpler; it does not share the drawbacks mentioned in the last paragraph. In other words the bias of the Gasser–Müller estimator is superior to that of the Nadaraya–Watson estimator.

Can one find an estimator that overcomes the disadvantages of these kernel methods? We introduce in Section 2 a design-adaptive regression method based on a weighted local linear regression that repairs the drawbacks of the two popular kernel smoothers (see Table 1). It will be shown that such a method adapts to various design densities, to both fixed and random designs, and to both interior and boundary points. Because of these adaptations, we sometimes refer to it as a design-adaptive regression estimator.

To gain an intuition on the benefits of the local linear regression smoother, consider Figures 1 and 2. Here the optimal bandwidths are used in computing pointwise asymptotic mean squared errors (MSE's). In Figure 1.a, the MSE for the local linear regression smoother is the smallest, and the Nadaraya–Watson estimator performs better than the Gasser–Müller estimator in a region close to 0, and performs worse in the other region. The performance of the Nadaraya–Watson estimator worsens in the case $\lambda = \frac{1}{4}$ (the quantity $f'_X(x)/f_X(x) = 1/\lambda$ is twice as large as the first case). It can be seen that the performance of the Nadaraya–Watson estimator worsens as x increases, because $|m'(x)|$ becomes larger. A similar phenomenon is illustrated in Figure 2. The reason that the Nadaraya–Watson estimators worsens for large $|x|$ is that $|f'_X(x)/f_X(x)|$ increases. It will be shown that the Gasser–Müller estimator has an asymptotic relative efficiency of 66.7%. Thus, the relative efficiency of the Nadaraya–Watson estimator is much less than 66.7% in Figure 1.b and 2.b.

The local linear smoother not only is superior to the two popular kernel regression estimators, but also is the best among all linear smoothers, including those produced by

* Jianqing Fan is Assistant Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260. This work was partially supported by NSF Grants DMS-9005905 and DMS-9113527. The author thanks an associate editor and two referees for their constructive and helpful comments.

Journal of the American Statistical Association
December 1992, Vol. 87, No. 420, Theory and Methods

Table 1. Pointwise Bias and Variance of Kernel Regression Smoothers

Method	*Bias*	*Variance*
Nadaraya–Watson	$\left(\frac{1}{2}m''(x) + \frac{m'(x)f_x'(x)}{f_x(x)}\right)\int_{-\infty}^{\infty} u^2K(u)\,du h_n^2$	$\frac{\sigma^2(x)}{f_x(x)nh_n}\int_{-\infty}^{\infty} K^2(u)\,du$
Gasser–Müller	$\frac{1}{2}m''(x)\int_{-\infty}^{\infty} u^2K(u)\,du h_n^2$	$\frac{3\sigma^2(x)}{2f_x(x)nh_n}\int_{-\infty}^{\infty} K^2(u)\,du$
Local linear smoother	$\frac{1}{2}m''(x)\int_{-\infty}^{\infty} u^2K(u)\,du h_n^2$	$\frac{\sigma^2(x)}{f_x(x)nh_n}\int_{-\infty}^{\infty} K^2(u)\,du$

orthogonal series and spline methods. It will be shown in Section 4 that the best local linear regression smoother has 100% efficiency among all linear smoothers in a minimax sense. Moreover, Fan (in press) showed that it has a high minimax efficiency among *all possible estimators,* including nonlinear smoothers such as median regression. See also

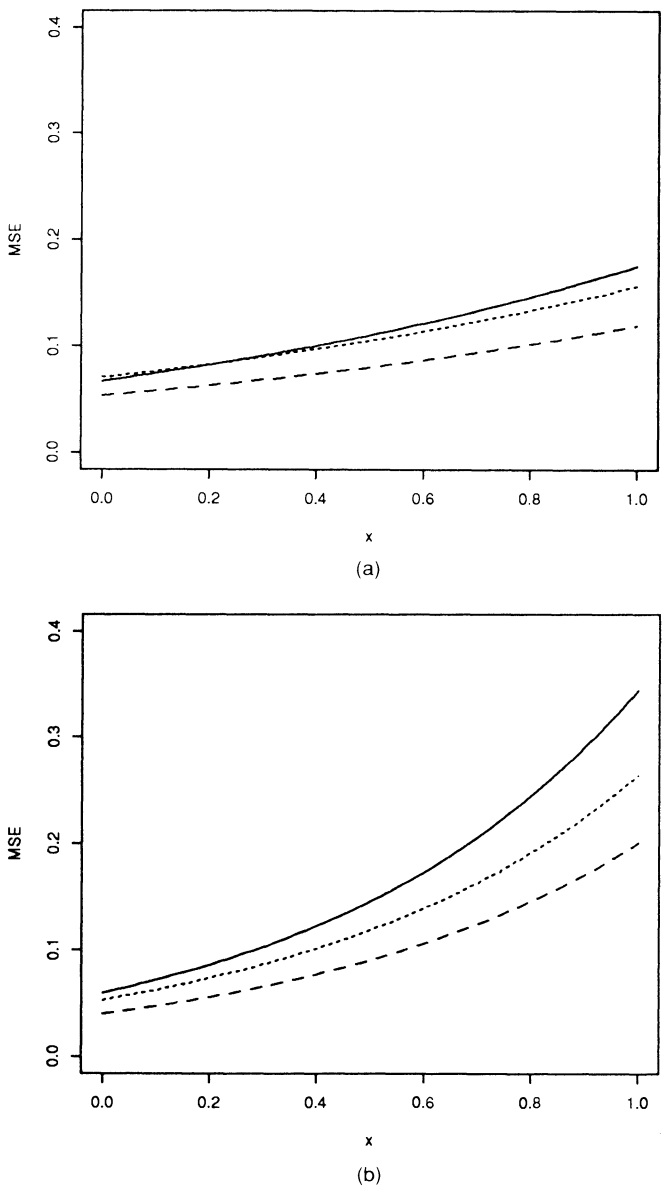


Figure 1. The Square Root of Pointwise Asymptotic MSE for the Regression Model $Y = -X^2 - 2X + .5\varepsilon$ With Sample Size $n = 100$ and the Covariate X Distributed Exponentially With Mean (a) $\lambda = \frac{1}{2}$ and (b) $\lambda = \frac{1}{4}$. The solid curve represents the MSE of the Nadaraya–Watson estimator, the dotted curve represents the MSE of the Gasser–Müller estimator, and the dashed curve represents the MSE of the local linear regression smoother.

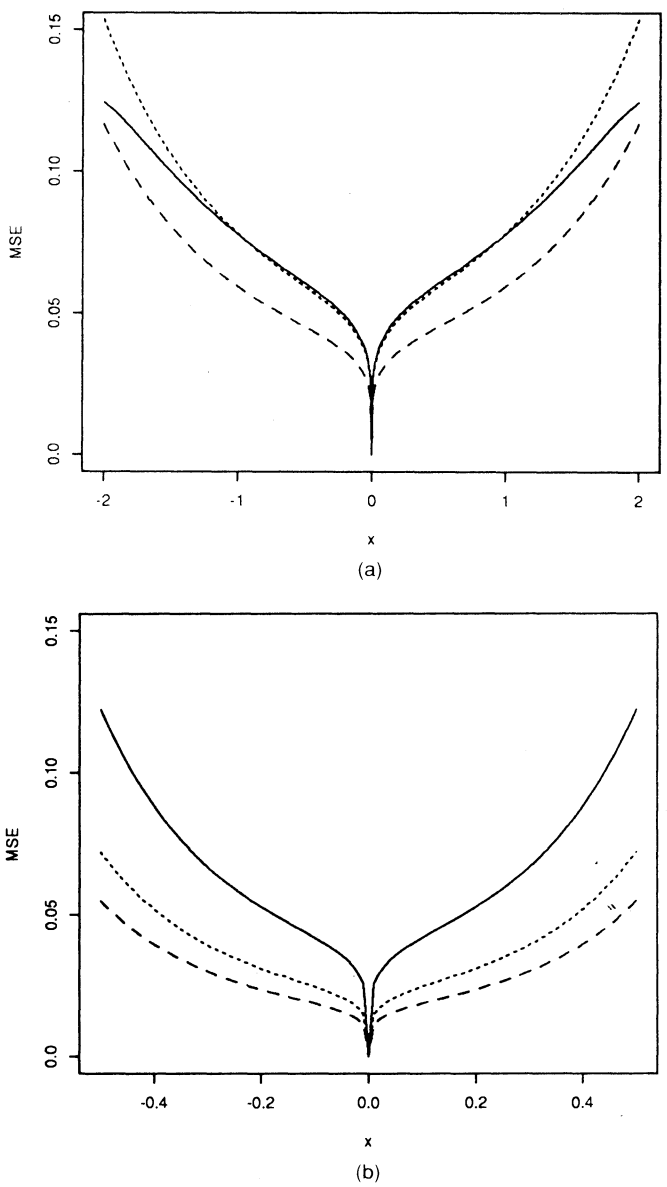


Figure 2. The Square Root of Pointwise Asymptotic MSE for the Regression Model $Y = \sin(3x/4) + .3\varepsilon$ With Sample Size $n = 100$ and the Covariate X Distributed as (a) $X \sim N(0, 1)$ and (b) $X \sim N(0, .25^2)$. The solid curve represents the MSE of the Nadaraya–Watson estimator, the dotted curve represents the MSE of the Gasser–Müller estimator, and the dashed curve represents the MSE of the local linear regression smoother.

Donoho and Liu (1991) for more discussion on minimax theory.

We illustrate the finite sample behavior of the local linear regression smoother via simulation studies. It turns out in Section 5 that the two popular kernel smoothers suffer low relative efficiency. Even in the uniform design case, one can still gain by using the local linear regression smoother, even though its asymptotic result is the same as the Nadaraya–Watson estimator.

2. LOCAL LINEAR REGRESSION

Suppose that the second derivative of $m(x)$ exists. In a small neighborhood of a point x, $m(y) \approx m(x) + m'(x)(y - x) \equiv a + b(y - x)$. Thus the problem of estimating $m(x)$ is equivalent to a local linear regression problem: estimating the intercept a. Now consider a weighted (local) linear regression: finding a and b to minimize

$$\sum_1^n (Y_j - a - b(X_j - x))^2 K\left(\frac{x - X_j}{h_n}\right). \tag{2.1}$$

Let $\hat{a}$ and $\hat{b}$ be the solution to the weighted least squares problem (2.1). Simple calculation yields

$$\hat{a} = \sum_1^n w_j Y_j \Big/ \sum_1^n w_j,$$

where w_j is defined by (2.3). Thus we define the local linear regression smoother by

$$\hat{m}(x) = \sum_1^n w_j Y_j \Big/ \sum_1^n w_j \tag{2.2}$$

with

$$w_j \equiv K\left(\frac{x - X_j}{h_n}\right)[s_{n,2} - (x - X_j)s_{n,1}], \tag{2.3}$$

where

$$s_{n,l} = \sum_1^n K\left(\frac{x - X_j}{h_n}\right)(x - X_j)^l, \qquad l = 1, 2. \tag{2.4}$$

This idea is an extension of Stone (1977), who used a kernel function $K(x) = \frac{1}{2} 1_{[|x| \le 1]}$, and was studied by Cleveland (1979), Fan (in press), Lejeune (1985), Müller (1987), and Tsybakov (1986). Note that $\hat{m}(x)$ is a weighted average of the responses and is called a linear smoother in the literature. The intuition at the beginning of this section suggests that $\hat{b}$ estimates $m'(x)$. Discussions on the behavior of $\hat{b}$ are beyond the scope of this article.

The bandwidth h_n can be chosen either subjectively by data analysts or objectively by data. A frequently used bandwidth selection technique is the cross-validation method (Stone 1977), which chooses h_n to minimize

$$\sum_{j=1}^n (Y_j - \hat{m}_{-j}(X_j))^2,$$

where $\hat{m}_{-j}(\cdot)$ is the regression estimator (2.2), without using the jth observation (X_j, Y_j). An alternative method is the plug-in approach (Hall, Sheather, Jones, and Marron 1991), which offers a faster rate of convergence in the density estimation setting.

3. ASYMPTOTIC PROPERTIES

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a population (X, Y) with regression function $m(x) = E(Y|X = x)$. We need the following conditions:

1. The regression function $m(x)$ has a bounded and continuous second derivative.
2. The conditional variance $\sigma^2(x) = \mathrm{var}(Y|X = x)$ is bounded and continuous.
3. The marginal density f_X of the covariate X is continuous and bounded away from the zero in an interval (a_0, b_0).
4. The kernel function K is a bounded density function with $\int_{-\infty}^{\infty} xK(x)\,dx = 0$ and $\int_{-\infty}^{\infty} x^4 K(x)\,dx < \infty$.

In the sequel we always denote

$$c_K = \int_{-\infty}^{\infty} u^2 K(u)\,du, \qquad d_K = \int_{-\infty}^{\infty} K^2(u)\,du.$$

We state the following pointwise and global properties of the local linear regression smoother and omit their proofs.

Theorem 1. Under Conditions 1–4, if $h_n \to 0$ and $nh_n \to \infty$, then for $x \in (a_0, b_0)$ estimator (2.2) has the conditional MSE

$$E[(\hat{m}(x) - m(x))^2 | X_1, \ldots, X_n] = \frac{1}{4}(c_K m''(x))^2 h_n^4 + \frac{d_K}{nh_n}\frac{\sigma^2(x)}{f_X(x)} + o_P\left(h_n^4 + \frac{1}{nh_n}\right). \tag{3.1}$$

Theorem 2. Under the conditions of Theorem 1, the weighted MISE is given by

$$E\left[\int_{-\infty}^{\infty} (\hat{m}(x) - m(x))^2 w(x)\,dx \,\middle|\, X_1, \ldots, X_n\right] = \frac{c_K^2}{4}\int_{-\infty}^{\infty} (m''(x))^2 w(x)\,dx h_n^4 + \frac{d_K}{nh_n}\int_{-\infty}^{\infty}\frac{\sigma^2(x)}{f_X(x)} w(x)\,dx + o_P\left(h_n^4 + \frac{1}{nh_n}\right), \tag{3.2}$$

provided that the weight function has a support containing in (a_0, b_0).

Simple algebra yields the optimal bandwidth for the conditional mean integrated square error (MISE) (3.2) is

$$h_{\mathrm{opt}} = \left(\frac{d_K \int_{-\infty}^{\infty} f^{-1}(x)\sigma^2(x)w(x)\,dx}{c_K^2 \int_{-\infty}^{\infty} (m''(x))^2 w(x)\,dx}\right)^{1/5} n^{-1/5}.$$

These results are stated for the random design; they are essentially the same for the fixed design of form $X_j = G^{-1}(i/n) + o(1/n)$, where the function G is the cdf of f_X. Thus the results are also applicable to such a design, namely, the local linear regression smoother adapts to both fixed and random designs. It also adapts to both highly clustered and nearly uniform designs, as suggested by Theorems 1 and 2.

We remark that the unconditional versions of Theorems 1 and 2 remain valid. Indeed, with a slight technical modification—defining $\hat{m}^*(x) = \sum_{j=1}^n w_j Y_j / (\sum_{j=1}^n w_j + n^{-2})$ to

guard against zero denominator—Fan (in press) showed that

$$E(\hat{m}^*(x) - m(x))^2 = \frac{1}{4}(c_K m''(x))^2 h_n^4 + \frac{d_K}{nh_n}\frac{\sigma^2(x)}{f_X(x)} + o\left(h_n^4 + \frac{1}{nh_n}\right);$$

a similar result holds for the unconditional MISE.

When the design density f_X has a bounded support, say (0, 1), the performance of regression smoothers at boundary points usually differs from the performance at interior points. Theoretically, the rate of convergence at boundary points is slower. For example, the Watson–Nadaraya and Gasser–Müller estimators have boundary effects—bias of order $O(h_n)$ instead of $O(h_n^2)$—and require boundary modifications (see Müller 1988). But the local linear smoother (2.2) *does not require* such a modification. Consider, for example, the left boundary points $x_n = ch_n$ with a positive constant c. It can be shown (Fan and Gijbels 1992) that

$$E((\hat{m}(x_n) - m(x_n))^2 | X_1, \ldots, X_n) = \frac{1}{4}(\alpha_K(c) m''(0+))^2 h_n^4 + \frac{\beta_K(c)}{nh_n}\frac{\sigma^2(0+)}{f_X(0+)} + o_P\left(h_n^4 + \frac{1}{nh_n}\right),$$

where with $s_{l,c} = \int_{-\infty}^{c} K(u)u^l\,du$ $(l = 0, 1, 2, 3)$,

$$\alpha_K(c) = \frac{s_{2,c}^2 - s_{1,c}s_{3,c}}{s_{2,c}s_{0,c} - s_{1,c}^2},$$

$$\beta_K(c) = \frac{\int_{-\infty}^{c}(s_{2,c} - us_{1,c})^2 K^2(u)\,du}{[s_{2,c}s_{0,c} - s_{1,c}^2]^2}.$$

Note that this conditional MSE shares the same form as (3.1) and an interior point corresponds to $c = \infty$, namely $\alpha_K(\infty) = c_K$ and $\beta_K(\infty) = d_K$. Figure 3 plots the functions α_K^2 and β_K for the Gaussian kernel. Remark that $\alpha_K^2(c) \le \alpha_K^2(\infty)$ and $\beta_K(c) \ge \beta_K(\infty)$—the bias at the boundary is smaller than at the interior because the local linear approximation is used on a smaller interval, and the variance is larger at the boundary because of fewer data points. Thus estimator (2.2) has desired biases and variances at boundary points and does not require any modifications. Figure 3 also suggests that a point $1.5h_n$ away from the boundary can be considered to be an interior point when Gaussian kernel is employed, because the function β_K becomes flat when $c \ge 1.5$.

Remark 1. The pointwise MSE (3.1) holds uniformly in the class of joint densities

$$\mathcal{C}_2 = \Big\{ f(\cdot,\cdot) : |m''(y) - m(x) - m'(x)(y-x)| \le \frac{C}{2}(y-x)^2 \text{ with } f_X \text{ and } \sigma^2(x) \text{ independent of } m(x) \text{ and satisfying conditions 1–3} \Big\}. \quad (3.3)$$

Here the marginal density f_X and the conditional variance $\sigma^2(x)$ remain the same over $\mathcal{C}_2$, whereas the regression function $m(x)$ varies. The condition on the regression function is slightly weaker than $|m''(\cdot)| \le C$. In the next section this class of joint densities will be used to study the best linear smoother.

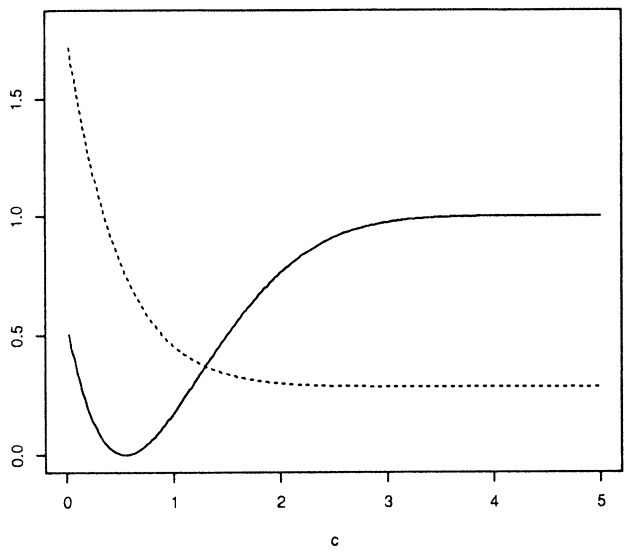


Figure 3. Constant Factors of the Asymptotic MSE at Boundary Points for the Gaussian Kernel. The solid curve represents the function α_K^2, and the dotted curve represents the function β_K.

Remark 2. The bias of the Nadaraya–Watson estimator

$$\hat{m}_{NW}(x) = \left[\sum_{j=1}^{n} K\left(\frac{x - X_j}{h_n}\right)Y_j\right]\left[\sum_{j=1}^{n} K\left(\frac{x - X_j}{h_n}\right)\right]^{-1} \quad (3.4)$$

depends on the derivatives of f_X. Hence its maximum risk over $\mathcal{C}_2$, say, is infinity and its asymptotic minimax efficiency is 0. But in the case of uniform designs, the Nadaraya–Watson estimator has the same asymptotic properties as the local linear regression smoother.

Remark 3. Gasser and Müller (1979) defined the following estimator:

$$\hat{m}_{GM}(x) = \sum_{j=1}^{n} Y_j' \int_{t_{j-1}}^{t_j} \frac{1}{h_n} K\left(\frac{x - t}{h_n}\right) dt, \quad (3.5)$$

where $\{(X_j', Y_j')\}$ are ordered samples according to X_j's, $t_0 = -\infty$, $t_n = \infty$, and $t_j = (X_j' + X_{j+1}')/2$. The variance of the local linear smoother is only two thirds of a corresponding Gasser–Müller estimator, while the bias is the same. (See Chu and Marron 1990; Jennen-Steinmetz and Gasser 1988; and Mack and Müller 1989 for the expression of the variance of the Gasser–Müller estimator.) Thus, in the case of random designs the latter estimator uses only two-thirds of the available data and is not admissible. For fixed designs, however, these two smoothers have the same asymptotic performance.

4. BEST LINEAR SMOOTHERS

Definition 1. A linear smoother is defined by the following weighted average:

$$\hat{m}_L(x) = \sum_{j=1}^{n} W_j(X_1, \ldots, X_n)Y_j. \quad (4.1)$$

It is obvious that estimator (2.2) is a linear smoother, as are the Nadaraya–Watson and Gasser–Müller estimators.

Estimation of regression using techniques of splines and orthogonal series is also linear. Thus linear smoothers include the most of regression techniques in the literature. For such a broad class of estimators, which one is the best? The answer depends on the class of regression function under consideration. For $\mathcal{C}_2$, the best design-adaptive regression estimator (defined in Theorem 3) is a best linear smoother.

Lemma 1. Let $\mathbf{a} = (a_1, \ldots, a_n)'$ and $\mathbf{w} = (w_1, \ldots, w_n)'$ be n-dimensional real vectors. Then

$$\min_{\mathbf{w}} \left[(\mathbf{w}'\mathbf{a} - b)^2 + \sum_{j=1}^{n} c_j w_j^2 \right] = \frac{b^2}{1 + \sum_{j=1}^{n} a_j^2/c_j},$$

and the minimizer is attained at

$$w_j = \frac{b}{1 + \sum_{j=1}^{n} a_j^2/c_j} a_j/c_j.$$

We omit the proof of Lemma 1. The conditional risk of linear smoother (4.1) is given by

$$\begin{aligned} R_1(m, \hat{m}_L) &\equiv E[(\hat{m}_L(x) - m(x))^2 | X_1, \ldots, X_n] \\ &= \left[\sum_{j=1}^{n} W_j m(X_j) - m(x) \right]^2 + \sum_{j=1}^{n} W_j^2 \sigma^2(X_j) \\ &\geq \frac{m^2(x)}{1 + \sum_{j=1}^{n} m^2(X_j)/\sigma^2(X_j)}. \end{aligned} \tag{4.2}$$

The latter inequality follows from Lemma 1. Note that the lower bound (4.2) is possessed by all linear smoothers, that is, it is independent of the smoother $\hat{m}_L$. Further study on (4.2) yields the following theorem.

Theorem 3. Let the minimax risk of linear smoothers be

$$R_L(n, \mathcal{C}_2) = \inf_{\hat{m}_L \text{linear}} \sup_{f \in \mathcal{C}_2} E([\hat{m}_L(x) - m(x)]^2 | X_1, \ldots, X_n),$$

where $\mathcal{C}_2$ was defined by (3.3). Assume that $\sigma^2(\cdot)$ is bounded away from 0 and infinity and the marginal density is bounded and continuous. Then

$$R_L(n, \mathcal{C}_2) = \frac{3}{4} 15^{-1/5} \left(\frac{\sqrt{C}\sigma^2(x)}{n f_X(x)} \right)^{4/5} (1 + o_P(1)), \tag{4.3}$$

and the best linear smoother $\hat{m}_0$ is given by (2.2) with kernel

$$K_0 = \frac{3}{4}[1 - x^2]_+ \quad \text{and} \quad h_n = \left(\frac{15\sigma^2(x)}{f_X(x) C^2 n} \right)^{1/5}$$

where C is defined by (3.3).

The quantity (4.3) plays a role similar to that of the Fisher information in the parametric inference. For a linear smoother $\hat{m}_L$, its minimax efficiency (among all linear smoothers) can be defined by

$$\text{efficiency of } \hat{m}_L = \left(\frac{R_L(n, \mathcal{C}_2)}{\sup_{f \in \mathcal{C}_2} E[(\hat{m}_L(x) - m(x))^2 | X_1, \ldots, X_n]} \right)^{5/4}. \tag{4.4}$$

By this definition an 80% efficient estimator uses only about 80% of the available data; the estimator with sample size 100 works as well as the optimal estimator with sample size 80.

Table 2. Minimax Efficiency for Kernel Regression Estimators

Kernel	Local linear smoother (%)	Gasser–Müller (%)	Nadaraya–Watson
Epanechnikov	100	66.67	0
Normal	95.12	63.41	0
Uniform	92.95	61.97	0

It can easily be calculated from (3.1) that the local linear regression smoother (2.2) has the minimax efficiency (with the bandwidth chosen to minimize its risk) of

$$.268 \left(\int_{-\infty}^{\infty} u^2 K(u)\, du \left(\int_{-\infty}^{\infty} K^2(u)\, du \right)^2 \right)^{-1/2},$$

where the factor .268 is computed from $(3/5 \times 15^{1/5})^{5/4}$. For random designs, the Gasser–Müller estimator (3.5) has the efficiency

$$\frac{2}{3} \times .268 \left(\int_{-\infty}^{\infty} u^2 K(u)\, du \left(\int_{-\infty}^{\infty} K^2(u)\, du \right)^2 \right)^{-1/2}.$$

The Nadaraya–Watson estimator has minimax 0, as indicated in Remark 2. Table 2 compares the minimax efficiencies of the three different types of kernel estimators for some commonly used kernels.

5. SIMULATIONS

We have shown, via asymptotics, that the local linear regression smoother possesses a number of desired properties. Nevertheless, its finite sample behavior is unknown. In this section we use three simulated examples to illustrate its finite sample behavior.

Simulation 1. A random sample of size n is simulated from the model

$$Y = \sin(.75X) + .3\varepsilon, \tag{5.1}$$

with $\varepsilon \sim N(0, 1)$ independent of $X \sim N(0, \sigma^2)$. When σ is small, the quantity $|f'_X(x)/f_X(x)|$ gets large. Thus we anticipate that the Nadaraya–Watson estimator does not behave well for small σ.

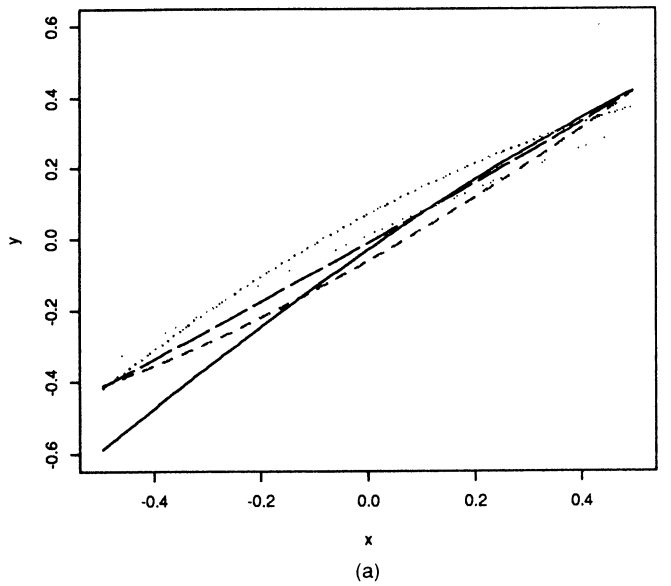
We estimate the regression function in the interval $x \in [-2\sigma, 2\sigma]$, two standard deviations away from its normal mean. The integrated square error (ISE) is computed by

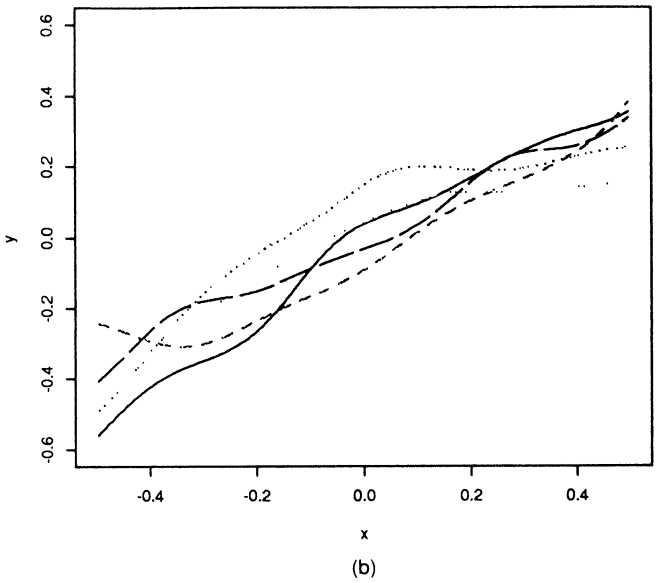
$$\text{ISE} = \int_{-2\sigma}^{2\sigma} (\hat{m}(x) - m(x))^2\, dx,$$

with an equal weight. The bandwidths of regression smoothers are chosen to minimize their asymptotic MISE, which

Table 3. MISE and Relative Efficiency Based on 300 Simulations

			Gasser–Müller		Nadaraya–Watson	
σ	Sample size	Local linear smoother MISE	MISE	Efficiency (%)	MISE	Efficiency (%)
.25	100	.0024	.0110	14.9	.0045	45.6
	200	.0013	.0066	13.1	.0028	38.3
	400	.0006	.0036	11.6	.0016	29.3
1.0	100	.0232	.0293	74.7	.0317	67.7
	200	.0122	.0166	68.0	.0171	65.6
	400	.0071	.0098	66.8	.0100	65.2





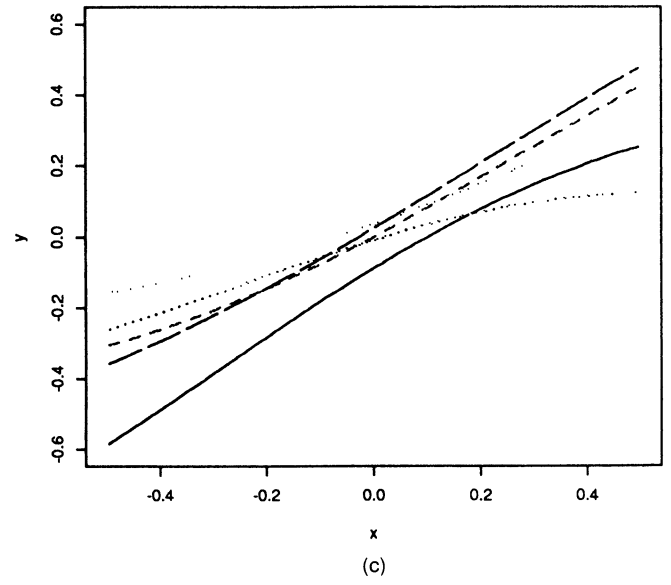


Figure 4. Estimates Based on Five Simulations from Model (5.1) With $n = 100$ and $X \sim N(0, .25^2)$: (a) Local linear regression smoother, (b) Nadaraya–Watson method, (c) Gasser–Müller approach.

can be computed from Table 1. Gaussian kernel function is used.

Table 3 shows the simulation results for the three different regression smoothers based on 300 simulations with $\sigma = .25$ and $\sigma = 1$. We also report the relative efficiency in comparison with the local linear regression smoother. The relative efficiency is computed by the following [compare with (4.4)]:

$$\text{relative efficiency of an estimator} = \left(\frac{\text{MISE of estimator (2.2)}}{\text{MISE of the estimator}}\right)^{5/4}$$

The relative efficiencies for both the Nadaraya–Watson and Gasser–Müller estimators are small. The Nadaraya–Watson estimator has smaller relative efficiencies for the case $\sigma = .25$ than it has for the case $\sigma = 1$. This seems compatible with the claim that the Nadaraya–Watson estimator can not be adaptive to highly clustered designs.

To visualize the global performance of the regression estimators, Figure 4 plots the estimates based on five repetitions for the case $n = 100$, $\sigma = .25$. In this case the regression curve is almost linear, and the Nadaraya–Watson estimate does not perform well in detecting linearity. Both the local linear regression smoother and the Gasser–Müller estimator are suitable for detecting linearity. But the variability of the Gasser–Müller estimator from one simulation to the other (i.e., variance) is larger.

Simulation 2. Instead of detecting linearity, consider the following model:

$$Y = \sin(2.5X) + .4\varepsilon, \tag{5.2}$$

with $\varepsilon \sim N(0, 1)$ and $X \sim .5N(-1, 1) + .5N(1.75, .25)$. The estimating procedures and computation are similar to those for Simulation 1. The ISE is computed by

$$\text{ISE} = \int_{-2}^{2} (\hat{m}(x) - m(x))^2 \, dx \tag{5.3}$$

for an estimator $\hat{m}(x)$. Table 4 reports the MISE and relative efficiencies.

Simulation 3. Instead of using the mixture normal density for the covariate X, we use Model (5.2) with marginal density X uniformly distributed on $[-2.5, 2.5]$ and compute the ISE defined by (5.3) based on 300 simulations. Table 5 reports the simulation results as well as the relative efficiencies.

Asymptotically, the performance of the local linear regression smoother is the same as the Nadaraya–Watson estimator. However, Table 5 indicates that the former smoother performs better than the latter at finite sample

Table 4. MISE and Relative Efficiency Based on 300 Simulations

Sample size	Local linear smoother MISE	Gasser–Müller MISE	Gasser–Müller Efficiency (%)	Nadaraya–Watson MISE	Nadaraya–Watson Efficiency (%)
100	0.1321	.1561	81.2	.1862	65.1
200	0.0701	.0868	76.6	.1009	63.4
400	0.0370	.0504	68.0	.0540	62.3

Table 5. MISE and Relative Efficiency Based on 300 Simulations

Sample size	Local linear smoother MISE	Gasser–Müller MISE	Gasser–Müller Efficiency (%)	Nadaraya–Watson MISE	Nadaraya–Watson Efficiency (%)
100	.0751	.0927	76.9	.0871	83.1
200	.0424	.0552	71.9	.0483	85.0
400	.0237	.0318	69.2	.0259	89.5

sizes. This suggests that the asymptotic theory takes in effect at a larger sample size for the Nadaraya–Watson estimator.

The efficiency of the Gasser–Müller estimator decreases when n increases. When $n = 400$, the relative efficiency is 69.2%, which is close to the asymptotic relative efficiency 66.7%.

APPENDIX: PROOF OF THEOREM 3

Because the sequence $X_1, \ldots, X_n$ are iid, it follows that

$$\sum_{j=1}^{n} m^2(X_j)/\sigma^2(X_j) = nEm^2(X_1)/\sigma^2(X_1) + O_p(\sqrt{nEm^4(X_1)/\sigma^4(X_1)}).$$

This, together with (4.2), leads to

$$R_1(m, \hat{m}_L) \geq \frac{m^2(x)}{1 + nEm^2(X_1)/\sigma^2(X_1) + O_p(\sqrt{nEm^4(X_1)/\sigma^4(X_1)})}. \quad \text{(A.1)}$$

Specifically, take $m_0(y) = (b_n^2/2)[1 - C(y - x)^2/b_n^2]_+$ so that $m_0 \in \mathcal{C}_2$, where $b_n = (15\sqrt{C}\,\sigma^2(x)/nf_X(x))^{1/5}$ maximizes (A.3). Then

$$\begin{aligned} Em_0^2(X_1)/\sigma^2(X_1) &= \frac{b_n^4}{4}\int_{-\infty}^{\infty}[1 - C(y-x)^2/b_n^2]_+^2 f_X(y)/\sigma^2(y)\,dy \\ &= \frac{b_n^5}{4}\int_{-\infty}^{\infty}[1 - Cz^2]_+^2 f_X(x + b_n z)/\sigma^2(x + b_n z)\,dz \\ &= \frac{b_n^5 f_X(x)}{4\sigma^2(x)}\int_{-\infty}^{\infty}[1 - Cz^2]_+^2\,dz(1 + o(1)) \\ &= \frac{4f_X(x)b_n^5}{15\sqrt{C}\sigma^2(x)}(1 + o(1)), \end{aligned} \quad \text{(A.2)}$$

and $Em^4(X_1)/\sigma^4(X_1) = O(b_n^8)$. By (A.1) and (A.2), we have

$$\begin{aligned} R_1(m_0, \hat{m}_L) &\geq \frac{b_n^4/4}{1 + \dfrac{4f_X(x)nb_n^5}{15\sqrt{C}\sigma^2(x)}(1 + o_P(1))} \\ &= \frac{3}{4}\,15^{-1/5}\left(\frac{\sqrt{C}\sigma^2(x)}{nf_X(x)}\right)^{4/5}(1 + o_P(1)). \end{aligned} \quad \text{(A.3)}$$

Thus

$$R_L(n, \mathcal{C}_2) \geq \frac{3}{4}\,15^{-1/5}\left(\frac{\sqrt{C}\,\sigma^2(x)}{nf_X(x)}\right)^{4/5}(1 + o_P(1)). \quad \text{(A.4)}$$

On the other hand, with $\hat{m}_0(x)$ defined by Theorem 3, by (3.1) one can easily show that

$$\begin{aligned} R_L(n, \mathcal{C}_2) &\leq \sup_{\mathcal{C}_2} R_1(m, \hat{m}_0) \\ &\leq \frac{3}{4}\,15^{-1/5}\left(\frac{\sqrt{C}\sigma^2(x)}{nf_X(x)}\right)^{4/5}(1 + o_P(1)). \end{aligned} \quad \text{(A.5)}$$

The result follows from (A.4) and (A.5).

[*Received February 1991. Revised September 1991.*]

REFERENCES

Chu, C. K., and Marron, J. S. (1992), "Choosing a Kernel Regression Estimator," *Statistical Science,* 6, 404–436.

Cleveland, W. S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association,* 74, 829–836.

Donoho, D. L., and Liu, R. C. (1991), "Geometrizing Rate of Convergence III," *The Annals of Statistics,* 19, 668–701.

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

Fan, J. (in press), "Local Linear Regression Smoothers and Their Minimax Efficiency," *The Annals of Statistics,* 20.

Fan, J., and Gijbels, I. (in press), "Variable Bandwidth and Local Linear Regression Smoothers," *The Annals of Statistics,* 20.

Gasser, T., and Müller, H. G. (1979), "Kernel Estimation of Regression Functions," in *Smoothing Techniques for Curve Estimation,* Lecture Notes in Mathematics 757, eds. T. Gasser and M. Rosenblatt, Heidelberg: Springer-Verlag, pp. 23–68.

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

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, 263–270.

Jennen-Steinmetz, C., and Gasser, T. (1988), "A Unifying Approach to Nonparametric Regression Estimation," *Journal of the American Statistical Association,* 83, 1084–1089.

Lejeune, M. (1985), "Estimation Nonparamètrique Par Noyaux: Règression Polynomiale Mobile, *Revue de Statistiques Appliquées,* 33, 43–68.

Mack, Y. P., and Müller, H. G. (1989), "Convolution-Type Estimators for Nonparametric Regression Estimation," *Statistics and Probability Letters,* 7, 229–239.

Müller, H. G. (1987), "Weighted Local Regression and Kernel Methods for Nonparametric Curve Fitting," *Journal of the American Statistical Association,* 82, 231–238.

——— (1988), *Nonparametric Analysis of Longitudinal Data,* Berlin: Springer-Verlag.

Nadaraya, E. A. (1964), "On Estimating Regression," *Theory of Probability and Its Applications,* 9, 141–142.

Stone, C. J. (1977), "Consistent Nonparametric Regression," *The Annals of Statistics,* 5, 595–620.

Tsybakov, A. B. (1986), "Robust Reconstruction of Functions by the Local-Approximation Method," *Problems of Information Transmission,* 22, 133–146.

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