Local Linear Smoothers in Regression Function Estimation

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

Irène Gijbels [1]
Department of Mathematics
Limburgs Universitair Centrum
B-3590 Diepenbeek, Belgium

Abstract

A method based on local linear approximation is used to estimate the mean regression function. The proposed local linear smoother has several advantages in comparison with other linear smoothers. Motivated by this fact, we follow this approach to estimate more general functions, among which, conditional median and conditional quantile functions. A further generalization involves the estimation of high-dimensional regression functions. In particular, additive and two-term interaction nonparametric regression models are discussed. This allows us to deal with the problem of "curse of dimensionality". Algorithms are provided to fit these models.

1 Introduction

This paper deals with various problems in nonparametric regression function estimation. The proposed estimator is based on a local linear approximation to the function being estimated. Such an estimator, called a local linear smoother, is discussed. In particular, we examine the optimalities and applications of this intuitively appealing approach.

[1] Completed while visiting Department of Statistics, University of North Carolina, Chapel Hill.

Abbreviated title. Local Linear Smoothers

AMS 1980 subject classification. Primary 62G05. Secondary 62G20.

Key words and phrases. Additive models; boundary effects; conditional quantile functions; convex loss; design-adaptive methods; interaction models; local linear smoothers; robustness.

With squared loss, the local linear smoother estimates the mean regression function. This estimator has various advantages compared with other kernel-type estimators, such as the Nadaraya-Watson (1964) estimator, and the Gasser-Müller (1979) estimator. More precisely, it adapts to both random and fixed designs, and to various design densities such as highly clustered designs and nearly uniform designs. It turns out that the local linear smoother repairs the drawbacks of other kernel regression estimators. Moreover, with an optimal choice of the kernel and the bandwidth, the local linear smoother is the best linear smoother. The proposed method also applies to both continuous and discrete type of response data. This is illustrated by simulations on a Binary response, a Poisson and a Normal regression model. Unlike other kernel-type smoothers, the local linear smoother does not have boundary effects: the behavior of the estimator at the boundary of the support of the design density is the same as in the interior. This fact is justified by both asymptotics and simulations.

Inspired by the above optimalities, we apply the proposed approximation to estimate other conditional functions. An example is the conditional median function which is especially useful when the error distribution is either asymmetric or heavy-tailed. Another estimated function is the conditional quantile function, which is of interest for describing the variation of bivariate data, and for obtaining prediction intervals. Our general approach leads also to robust estimators — this is illustrated by a comparison between conditional median and mean function. Like in the classical theory, both estimators have the same asymptotic bias, but the median regression estimator tends to have a smaller asymptotic variance.

We also discuss the problem of choosing an optimal bandwidth, locally and globally. The advantage of using an optimal local bandwidth is addressed.

A further generalization consists of considering regression problems in higher dimensions. A typical problem in dealing with high-dimensional data is the "curse of dimensionality". In order to cope with this problem, additive regression models and two-term interaction

regression models are discussed. The latter models also allow us to examine the interaction between covariates. An algorithm for estimating the main effects and the interactions is introduced. This algorithm is inspired by the ACE algorithm (Breiman and Friedman (1985)), and results into robust estimators.

The paper is organized as follows. The local linear smoother, together with its optimalities and simulated examples, is discussed in Section 2. Section 3 deals with the generalization of estimating other functions. Higher dimensional regression problems are studied in Section 4. This paper does not have the intention to give a detailed study of the whole topic. The main goal is to give an overview of the optimality and the wide applicability of the proposed local linear smoothing method. Some challenging open problems are addressed in Section 5. The last section contains the proofs of the presented results.

2 Local Linear Smoothers

Given a random sample $(X_1, Y_1), \cdots, (X_n, Y_n)$ of the covariate and the response, it is common practice to explore the association between the covariate and the response via regression analysis: estimating $m(x) = E(Y|X = x)$—the best predictor in mean squared error. The proposed smoothing method consists of using weighted local linear regression, which has advantages over other linear smoothers.

2.1 Weighted Local Linear Regression

It is not possible to estimate the regression curve $m(x)$ without some smoothness conditions. The condition that $m(x)$ has a bounded second derivative is typically used in the literature. Under this condition, $m(y) \approx m(x) + m'(x)(y - x) \equiv a + b(y - x)$, in a neighborhood of the point x. Thus, the problem of estimating $m(x)$ is equivalent to estimating the intercept a locally.

Let K be a bounded probability density function with mean zero, and denote by h_n the

bandwidth. Consider a weighted local linear regression: i.e. find a and b that minimize

$$\sum_{j=1}^{n}(Y_j - a - b(x - X_j))^2 K\left(\frac{x - X_j}{h_n}\right).$$

Let $\hat{a}, \hat{b}$ be the solution to this weighted least squares problem. The proposed regression estimator is

$$\hat{m}(x) = \hat{a} = \sum_{j=1}^{n} w_j Y_j / \sum_{j=1}^{n} w_j, \tag{2.1}$$

with

$$w_j \equiv K\left(\frac{x - X_j}{h_n}\right)[s_{n,2} - (x - X_j)s_{n,1}], \tag{2.2}$$

where

$$s_{n,l} = \sum_{j=1}^{n} K\left(\frac{x - X_j}{h_n}\right)(x - X_j)^l, \qquad l = 0, 1, 2. \tag{2.3}$$

First, note that the resulting nonparametric regression estimator is a weighted average of the responses, and is called a *linear smoother* in the literature. Secondly, the estimator (2.1) is derived by using a *local linear* approximation to the regression function. Therefore, we refer to the estimator (2.1) as a *local linear smoother.*

The above idea is an extension of Stone (1977) who uses the kernel function $K(x) = \frac{1}{2}1_{[|x|\leq 1]}$, and is studied by Cleveland (1979), Müller (1987) and Fan (1990, 1991).

The bandwidth h_n can be determined by using the cross-validation method: find h_n that minimizes

$$\sum_{j=1}^{n}(Y_j - \hat{m}_{-j}(X_j))^2, \tag{2.4}$$

where $\hat{m}_{-j}(\cdot)$ is the regression estimator (2.1) without using the j^{th} observation (X_j, Y_j). See Stone (1977).

2.2 Asymptotic Properties

Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be a random sample from a population (X, Y) with regression function $m(x) = E(Y|X = x)$. Under some mild conditions, Fan (1991) showed that the conditional Mean Squared Error (MSE) of the local linear smoother is given by

$$E\left[(\hat{m}(x) - m(x))^2 \middle| X_1, \cdots, X_n\right] = b_n^2(x) + v_n^2(x) + o_P(h_n^4 + \frac{1}{nh_n}), \tag{2.5}$$

with

$$b_n(x) = \frac{1}{2}m''(x)\left(\int_{-\infty}^{\infty} v^2 K(v)dv\right)h_n^2,$$

and

$$v_n^2(x) = \frac{\sigma^2(x)}{nh_n f_X(x)}\int_{-\infty}^{\infty} K^2(v)dv, \tag{2.6}$$

where $f_X(x)$ is the marginal density of the covariate X, and $\sigma^2(x) = \text{var}(Y|X = x)$ is the conditional variance of the response given the covariate. A similar result holds for the unconditional MSE (see Fan (1990)).

The above result is obtained for random designs. Such design-models are suitable for cases where the covariates are beyond the experimenters' control, as is the case in for example height and weight studies. In many other cases, design points are prescribed by data analysts. In what follows, we extend the previous result to fixed designs.

Let $(x_1, Y_1), \cdots, (x_n, Y_n)$ be a random sample from the regression model

$$Y_j = m(x_j) + \sigma(x_j)\varepsilon_j, \tag{2.7}$$

where ε_j are i.i.d. random variables with mean zero and variance one. Note that this model allows for heteroscadecity. Let $f_X(x)$ be the design density, i.e.

$$x_j = G^{-1}(j/n) \quad \text{with} \quad G(x) = \int_{-\infty}^{x} f_X(y)dy. \tag{2.8}$$

Let $\hat{m}(x)$ be the estimator (2.1) with (X_j, Y_j) replaced by (x_j, Y_j).

Theorem 1. *Assume that $m(\cdot)$ has a bounded and continuous second derivative, $\sigma(\cdot)$ is uniformly Lipschitz continuous, and $f_X(\cdot)$ is continuous and bounded away from zero uniformly in its support $[a, b]$. Let K be a bounded density with mean 0. Assume further that $z^l K(z)$, for $l = 0, 1, 2$, is a bounded, and uniformly Lipschitz continuous function. Suppose that there exists a $\delta \in (0, 1)$ such that $nh_n^{1+3\delta} \to \infty$, and $\lim_{|x|\to\infty} x^{3+\delta}K(x^\delta) = 0$. Then the estimator (2.1) has mean squared error:*

$$E(\hat{m}(x) - m(x))^2 = b_n^2(x) + v_n^2(x) + o(h_n^4 + \frac{1}{nh_n}), \quad \text{for } x \in (a, b),$$

where b_n and v_n^2 are given by (2.6).

Hence, the estimator $\hat{m}(x)$ adapts to both random designs and fixed designs.

For the asymptotic normality result of the estimator in case of fixed designs, see Müller (1987).

In the situation of random designs, the asymptotic variance of the Gasser-Müller estimator is 1.5 times as large as that of the local linear smoother, while the asymptotic bias is the same. Thus, the Gasser-Müller estimator can not adapt to random designs. The estimator is also asymptotically inadmissible, with asymptotic relative efficiency 66.7% compared to the local linear smoother. See Fan (1991) for the efficiency calculation and Jennen-Steinmetz and Gasser (1988), Mack and Müller (1989), Chu and Marron (1990) for the expression of and a discussion on the asymptotic variance of the Gasser-Müller estimator.

In comparison with the Nadaraya-Watson estimator, the local linear smoother has a simpler asymptotic bias. It appears that the asymptotic bias of the Nadaraya-Watson estimator depends on $m'(x)f_X'(x)/f_X(x)$, which does not reflect an intrinsic difficulty of nonparametric regression, but an artifact produced by that estimation method. The Nadaraya-Watson estimator can not adapt to designs where $|f_X'(x)/f_X(x)|$ is large. Further, its bias can be arbitrarily large even in the situation when the true regression function is linear. Moreover, the minimax efficiency of the estimator is zero. In contrast, the local linear smoother adapts to designs where $|f_X'(x)/f_X(x)|$ is large.

For the reasons mentioned above, we also refer to the proposed method as a *design-adaptive method.*

Chu and Marron (1990) and Fan (1990, 1991) give detailed discussions on the three kernel approaches: the Nadaraya-Watson, the Gasser-Müller and the design-adaptive estimator. These studies include asymptotic efficiencies, numerical examples and finite sample comparisons. It turns out that the design-adaptive method is the best method. Moreover, the design-adaptive regression estimator is the best linear smoother (see Section 2.4). Because of its advantages, we would expect that it will become the benchmark of nonpara-

metric regression.

2.3 Boundary Effects

Consider the situation where the design density has a bounded support, say $[0,1]$. It is known that the performance of both the Nadaraya-Watson and the Gasser-Müller estimator at the boundary differs from that in the interior. Consider, as an example, left-boundary points $x_n = ch_n$ with $c > 0$. The rate of convergence of both estimators at those points is slower than that in the interior. In the literature, one refers to this problem as "boundary effects".

In the next theorem, we investigate the behavior of the local linear smoother (2.1) at the boundary of the support. To fix the idea we consider left-boundary points of the form $x_n = ch_n$. Further, for simplicity, we only deal with the setup of random designs. Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be a random sample from a population (X, Y), for which $f_X(\cdot)$ has support $[0,1]$.

The following theorem is a special case of Theorem 4 in Fan and Gijbels (1991).

Theorem 2. *Assume that $f_X(\cdot)$, $m''(\cdot)$ and $\sigma(\cdot)$ are all bounded on $[0,1]$, and right continuous at the point 0. Suppose that $\limsup_{v\to-\infty}|K(v)v^5| < \infty$. Then, the conditional MSE of the estimator (2.1) at the boundary point x_n is given by*

$$\left\{\frac{1}{4}\left[m''(0+)\frac{s_{2,c}^2 - s_{1,c}s_{3,c}}{s_{2,c}s_{0,c} - s_{1,c}^2}\right]^2 h_n^4 + \frac{\int_{-\infty}^{c}[s_{2,c} - vs_{1,c}]^2K^2(v)dv}{[s_{2,c}s_{0,c} - s_{1,c}^2]^2}\right\}\frac{\sigma^2(0+)}{f_X(0+)nh_n}(1 + o_P(1)), \tag{2.9}$$

where $s_{l,c} = \int_{-\infty}^{c} K(v)v^l dv, (l = 0,1,2,3)$.

A similar result holds for right-boundary points (see Remark 2 in Fan and Gijbels (1991)).

From the above results, it is clear that the local linear smoother performs equally well at the boundary as in the interior of the support. Indeed, the rate of convergence of the estimator does not depend on the location of the point under consideration. This is not surprisingly, taking into account the way the estimator is constructed. More precisely, the approximation $m(y) \approx m(x) + m'(x)(y-x)$ forces the error to be of order $(y-x)^2$, which behaves like $O(h_n^2)$ for effective design points. This approximation holds for both interior and boundary points.

As to our knowledge, this is the only kernel-type estimator without "boundary effects". This fact even more supports the proposed design-adaptive method and highlights one of its 'natural' features.

2.4 Asymptotic Optimalities

The design-adaptive regression method is intuitively appealing. It overcomes disadvantages of other kernel methods. One would expect that the method is the best linear method.

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

$$\hat{m}_L(x) = \sum_{j=1}^{n} W_j(X_1, \cdots, X_n) Y_j \tag{2.10}$$

Note that this definition does not assume that the total weight is one, which broadens the class of estimators. The linear smoothers include estimators produced by the method of kernel, orthogonal series, and splines. So, most regression smoothers in the literature are linear smoothers. The conditional median smoother however, is not a linear smoother.

Consider a class of joint densities, say $\mathcal{C}_1$, whose marginal density of X is independent of m. Furthermore, let

$$\mathcal{C}_2 = \{f(\cdot,\cdot) \in \mathcal{C}_1 : |m(y) - m(x) - m'(x)(y-x)| \leq C(y-x)^2, |m(x)| \leq D\}, \tag{2.11}$$

where C and D denote some positive constants. Here, we consider the class of joint densities whose conditional distribution $f(y|x)$ is parameterized by the mean regression function

$m(x)$. For this reason, we use $\sigma_m^2(x)$ to denote the conditional variance of Y given X. This variance can either be fixed or vary, depending on whether it is a function of $m(\cdot)$ or not. This class of densities is more general than the class considered by Fan (1991). In that paper it is assumed that the conditional variance is independent of $m(\cdot)$, which rules out the Binary response and the Poisson models discussed in Examples 1 and 2 below.

The following theorem shows that the design-adaptive nonparametric regression estimator is the best linear smoother. It is an extension of Theorem 3 in Fan (1991).

Theorem 3. *Assume that $f_X(\cdot)$ is continuous at the point x and $a \equiv \sup_{f \in \mathcal{C}_2} \sigma_m^2(x) < \infty$. Suppose that $\sigma_m^2(x)$ depends on $m(\cdot)$ only through the value $\theta \equiv m(x)$, and is a continuous function of θ. Then, the local linear smoother (2.1), with kernel $K(x) = \frac{3}{4}[1-x^2]_+$, and $h_n = \left(\frac{15a}{4f_X(x)C^2n}\right)^{1/5}$, denoted by $\hat{m}^*(x)$, is the best linear smoother in the sense that for any linear smoother $\hat{m}_L(x)$,*

$$\sup_{f \in \mathcal{C}_2} E\left[[\hat{m}_L(x) - m(x)]^2 \middle| X_1, \cdots, X_n\right] \geq r(1 + o_P(1))$$

and on the other hand $\sup_{f \in \mathcal{C}_2} E\left[[\hat{m}^*(x) - m(x)]^2 \middle| X_1, \cdots, X_n\right] = r(1 + o_P(1))$, where

$$r = \frac{3}{4}\left(\frac{15}{4}\right)^{-1/5}\left(\frac{\sqrt{C}a}{nf_X(x)}\right)^{4/5}. \tag{2.12}$$

The linear minimax risk r can be used to measure the efficiency of a regression estimator. For more discussions on minimax theory in density estimation and Gaussian white noise models, see Donoho and Liu (1991a,b).

Example 1. (*Binary response model*). Consider a random sample $(X_1, Y_1), \cdots, (X_n, Y_n)$ with

$$P\{Y_j = 1 | X_j\} = m(X_j), \qquad P\{Y_j = 0 | X_j\} = 1 - m(X_j).$$

Then, $\sigma_m^2(x) = m(x)[1 - m(x)]$. According to Theorem 3, the design-adaptive regression estimator (2.1) is the best linear smoother, with $a = D(1-D) \leq 1/4$ $(0 < D < 1)$. One

can use this nonparametric method as a diagnostic tool to see whether a more traditional logistic linear model is reasonable or not.

Example 2. (*Poisson regression*). Given the covariate $X = x$, the response variable Y has a Poisson distribution with mean $m(x)$. Then, the conditional variance $\sigma_m^2(x) = m(x)$. Theorem 3 assures that the design-adaptive regression estimator is the best linear smoother (with $a = D$).

Example 3. (*Normal regression*). Given the covariate $X = x$, $Y \sim N(m(x), \sigma^2(x))$. Here, the conditional variance is independent of $m(x)$, and the best linear smoother is given by Theorem 3, with $a = \sigma^2(x)$.

2.5 Simulations

In this section, we illustrate the applicability of the local linear smoother to both continuous and discrete models. The presented examples also demonstrate the absence of boundary effects. To eliminate confounding factors in studying these effects, we work with a uniform(0,1) random design density $f_X(\cdot)$. Further, we take a standard normal kernel function and use the optimal constant bandwidth

$$h_0 = \left(\frac{\int_0^1 \sigma^2(x)dx}{2\sqrt{\pi} \int_0^1 [m''(x)]^2 dx} \right)^{1/5} n^{-1/5},$$

which is a special case of expression (3.11).

First of all, we simulate from a Binary response model with

$$m(x) = (3 + 2\cos(16x) + \sin(10x))/6.$$

The result of 10 simulations (each of sample size 400), together with the true regression curve, is presented in Figure 1.

Example 2 concerns a discrete Poisson regression model with

$$m(x) = 14\exp(-(x - 0.5)^2).$$

Figure 2 shows 10 estimated curves, based on random samples of size 400.

In a third example, a Normal regression model with

$$m(x) = 3x^3 + \sin(4\pi x) \qquad \text{and} \qquad \sigma^2(x) = 0.5$$

is fitted. Simulations based on random samples of size 100 resulted into Figure 3.

Insert here Figures 1 — 3.

3 Design-adaptive Function Regression

In this section, we apply the design-adaptive regression procedure, discussed in Section 2, to a more general setting. We provide a *robust* and *unified approach* which allows data analysts to analyze various types of regression problems.

3.1 Generalized Local Linear Regression

We are interested in estimating the function $m_\rho(x)$ that minimizes

$$E(\rho(Y - m(x))|X = x), \tag{3.1}$$

with respect to $m(x)$, where $\rho(\cdot)$ is a convex function with a minimum at the point $x = 0$.

The estimation procedure is as follows: find $\hat{a}$ and $\hat{b}$ that minimize

$$\sum_{j=1}^{n} \rho\left(Y_j - a - b(x - X_j)\right) K\left(\frac{x - X_j}{h_n}\right), \tag{3.2}$$

with respect to a and b. We propose to use $\hat{m}_\rho(x) = \hat{a}$ in order to estimate $m_\rho(x)$. We refer to this estimator as a *generalized local linear smoother.*

Some applications of the above estimation procedure are:

- with $\rho(z) = |z|$—local least deviation regression, the resulting estimator is estimating the conditional median function; this is very useful when error distributions are either asymmetric or heavy-tailed.

- with $\rho(z) = pz_+ + (1-p)z_-$, we estimate the conditional p-quantile of Y given $X = x$; the estimated conditional p-quantile is very useful for obtaining a prediction interval of the response Y given the value of the covariate $X = x$;
- with $\rho(z) = z^2$—local least squares regression, the resulting estimator is estimating the conditional mean function proposed in (2.1);
- with $\rho'(z) = \psi_d(z)$, where

$$\psi_d(z) = z1_{(|z|\leq d)} + d1_{(z>d)} - d1_{(z<-d)} \tag{3.3}$$

is Huber's ψ-function (see Huber (1981)), we obtain a robust regression estimator.

Härdle and Gasser (1984) and Hall and Jones (1990) proposed to minimize (3.2) without the linear term $b(x - X_j)$. In other words, they consider a generalization of the Nadaraya-Watson estimator – and this generalization shares the disadvantages of that estimator.

3.2 Asymptotic Normality

In an unpublished manuscript by Fan, Hu and Truong (1991), it is shown that under some conditions,

$$\hat{m}_\rho(x) - m_\rho(x) \xrightarrow{\mathcal{L}} N(b_n(x), v_n^2(x))$$

where

$$b_n(x) = \frac{1}{2}m''_\rho(x)\left(\int_{-\infty}^{\infty} v^2 K(v)dv\right)h_n^2, \tag{3.4}$$

and

$$v_n^2(x) = \sigma_\rho^2(x)\frac{\int_{-\infty}^{\infty} K^2(v)dv}{nh_n f_X(x)} \tag{3.5}$$

with

$$\sigma_\rho^2(x) = \frac{\int_{-\infty}^{\infty}[\rho'(y - m_\rho(x))]^2 f(y|x)dy}{[\int_{-\infty}^{\infty}\rho''(y - m_\rho(x))f(y|x)dy]^2}, \tag{3.6}$$

where $f(y|x)$ is the conditional density of Y given $X = x$.

Note that the bias $b_n(x)$ does not depend on the function ρ in a direct way. It involves only the second derivative of the object function. The asymptotic variance however, depends

on ρ through σ_ρ. See also Tsybakov (1986) for related results, concerning a special choice of ρ.

3.3 Optimal Local Bandwidth

Denote by h_x the bandwidth which minimizes $b_n^2(x) + v_n^2(x)$. Then,

$$h_x = \left(\frac{\sigma_\rho^2(x) \int_{-\infty}^{\infty} K^2(v)dv}{f_X(x)[m_\rho''(x)]^2[\int_{-\infty}^{\infty} v^2 K(v)dv]^2} \right)^{1/5} n^{-1/5}. \tag{3.7}$$

With this optimal bandwidth, the Asymptotic Mean Squared Error (AMSE) is given by

$$b_n^2(x) + v_n^2(x) = \frac{5C_K}{4n^{4/5}} [m_\rho''(x)]^{2/5} \, [\sigma_\rho^2(x)/f_X(x)]^{4/5}, \tag{3.8}$$

where $C_K = [\int_{-\infty}^{\infty} v^2 K(v)dv]^{2/5} \, [\int_{-\infty}^{\infty} K^2(v)dv]^{4/5}$. Now, if we take as measure of global loss

$$\text{AMISE}(\hat{m}_\rho, m_\rho) = \int_{-\infty}^{\infty} (b_n^2(x) + v_n^2(x)) w(x)dx, \tag{3.9}$$

where $w(\cdot)$ is an integrable weight function, then with optimal local bandwidth h_x at each location, the estimator has

$$\text{AMISE}_1 = \frac{5C_K}{4n^{4/5}} \int_{-\infty}^{\infty} [m_\rho''(x)]^{2/5} \, [\sigma_\rho^2(x)/f_X(x)]^{4/5} w(x)dx. \tag{3.10}$$

In contrast, if one takes the optimal constant bandwidth

$$h_0 = \left(\frac{\int_{-\infty}^{\infty} \sigma_\rho^2(x) w(x)/f_X(x)dx \int_{-\infty}^{\infty} K^2(v)dv}{\int_{-\infty}^{\infty} [m_\rho''(x)]^2 w(x)dx \, [\int_{-\infty}^{\infty} v^2 K(v)dv]^2} \right)^{1/5} n^{-1/5} \tag{3.11}$$

that minimizes (3.9), then the best AMISE is given by

$$\text{AMISE}_2 = \frac{5C_K}{4n^{4/5}} \left(\int_{-\infty}^{\infty} [m_\rho''(x)]^2 w(x)dx \, [\int_{-\infty}^{\infty} \sigma_\rho^2(x)/f_X(x) w(x)dx]^4 \right)^{1/5}. \tag{3.12}$$

Comparing (3.10) with (3.12), it is easy to see that (3.10) is smaller, for example by using the Cauchy-Schwarz inequality. In other words, the estimator with optimal local bandwidth does a better job than the one with optimal constant bandwidth.

3.4 Comparison between mean and median regression

Let $f(\cdot|x)$ be symmetric about $m(x)$, i.e. let the error distribution be symmetric. If we take $\rho_d'(z) = \psi_d(z)$ given by (3.3), then the method (3.2) results in estimating the location function $m(x)$, which is both conditional mean and conditional median. With such a choice of ρ,

$$\sigma^2_{\rho_d}(x) = \frac{E[\psi_d^2(Y - m(x))|X = x]}{[P(|Y - m(x)| \leq d|X = x)]^2}.$$

Evidently, $\lim_{d\to 0} \sigma^2_{\rho_d}(x) = \frac{1}{4f^2[m(x)|x]}$. This suggests that the asymptotic variance of the median regression estimator is given by

$$\sigma^2_{med}(x) = \frac{1}{4f^2[m(x)|x]} \frac{\int_{-\infty}^{\infty} K^2(v)dv}{nh_n f_X(x)}. \tag{3.13}$$

Note that the asymptotic variance of the mean regression estimator is displayed in (2.6). Denote this asymptotic variance by $\sigma^2_{mean}(x)$. Thus, the mean and median regression estimator have the same asymptotic bias but different asymptotic variances. This result is very similar to the classical theory on the mean and median estimator. Here, the ratio of the asymptotic variances is given by

$$\frac{\sigma^2_{mean}(x)}{\sigma^2_{med}(x)} = 4f^2[m(x)|x]\sigma^2(x). \tag{3.14}$$

It follows from the well-known result on the comparison between the mean and median (see Theorem 3.3 of Lehmann (1983)) that the ratio in (3.14) is greater than or equal to $1/3$, provided that $f[m(x)|x] \geq f[z|x]$ for all z (a weakening condition of unimodality). Note that if the error distribution is Cauchy, the ratio in (3.14) is infinite, and if the error distribution is uniform, the lower bound $1/3$ is attained.

So, like in the classical theory, median regression is a more robust regression method than mean regression. Moreover, it is easier to interpret, especially in the case of asymmetric error distributions.

4 Generalized Additive Nonparametric Regression

Consider a regression problem with more than one covariate, i.e. let Y, $X_1, \cdots, X_p$ be random variables with Y the response and $X_1, \cdots, X_p$ the predictors. The association between the response and the covariates can be studied through the function $m_\rho(x_1, \cdots, x_p)$ that minimizes

$$E\left[\rho(Y - m(x_1, \cdots, x_p))|\, X_1 = x_1, \cdots, X_p = x_p\right]. \tag{4.1}$$

Such a function m_ρ is a multivariate extension of the function defined by (3.1). Applications of such a function were mentioned in Section 3.1.

When the number of predictors p is large – which is usually the case for many scientific studies – it is very difficult to explore the association between the response Y and the covariates $X_1, \cdots, X_p$ because of the intrinsic difficulty known as "curse of dimensionality". To cope with this difficulty, a low-dimensional least squares regression problem has been proposed (see Stone (1990) for interesting discussions on the optimal rates). In particular, a popular way to avoid the "curse of dimensionality" is to use an additive regression model. Such a model was suggested by Friedman and Stuetzle (1981) and forms the core for the ACE algorithm (Breiman and Friedman (1985)). See also Buja, Hastie and Tibshirani (1989), Donoho and Johnstone (1989), and Hall (1989). Estimation of m_ρ via the best additive approximation will be discussed in Section 4.1.

An interesting feature in analyzing high-dimensional data is to study the interactions between the covariates. However, the additive regression method is incapable of handling such a feature. To cope with the curse of dimensionality and to incorporate interactions, a two-term interaction nonparametric regression model is introduced in Section 4.2.

4.1 Additive Nonparametric Regression

Instead of estimating $m_\rho(x_1, \cdots, x_p)$ directly, one estimates $c, \phi_1(\cdot), \cdots, \phi_p(\cdot)$ that minimize the error not explained by the additive functions

$$E\rho\left(Y - c - \sum_1^p \phi_j(X_j)\right), \tag{4.2}$$

subject to the normalizations that

$$E\phi_j(X_j) = 0, \qquad j = 1, \cdots, p. \tag{4.3}$$

So, the additive method focuses on finding the best additive prediction, which can also be viewed as the best additive approximation to the object m_ρ defined by (4.1). In particular, with $\rho(z) = z^2$, the additive functions are the same as those which are estimated via the ACE algorithm. With $\rho'(z) = \psi_d(z)$, this method can be viewed as a robustification of the ACE method. For this reason, we refer to the following estimate as an *iterated robustness enhanced nonparametric estimate.*

Let $(y_j; x_{j1}, \cdots, x_{jp}), j = 1, \cdots, n$, be a set of observations from $(Y, X_1, \cdots, X_p)$. The proposed algorithm for estimating the additive functions is as follows:

step 0. Estimate the constant c by $\hat{c}$ that minimizes $\sum_{j=1}^n \rho(y_j - c)$. Let $r_j = y_j - \hat{c}$ be the residuals.

step 1. Regress $\{r_j, j = 1, \cdots, n\}$ on $\{x_{j1}, j = 1, \cdots, n\}$ by using the generalized local linear smoother (see (3.2)). The kernel function is taken to be the standard normal density and h_n is determined by the cross-validation criterion (see (2.4)). Denote this local linear smoother by $\hat{\phi}(\cdot)$. Put $\hat{\phi}_1^*(\cdot) = \hat{\phi}_1(\cdot) - \frac{1}{n}\sum_{j=1}^n \hat{\phi}_1(x_{j1})$, in order to cope with the constraint (4.3). Let $\{R_j = r_j - \hat{\phi}_1^*(x_{j1}), j = 1, \cdots, n\}$ be the resulting residuals.

step 2. Regress $\{R_j, j = 1, \cdots, n\}$ on $\{x_{j2}, j = 1, \cdots, n\}$ to estimate $\phi_2(\cdot)$ in the same manner as step 1.

step 3. Keep estimating the additive functions in the above manner until $\phi_p(\cdot)$ is estimated.

step 4. Keep cycling steps 0-3 until the mean of the residuals $\frac{1}{n}\sum_{j=1}^{n}\rho(R_j)$ fails to decrease dramatically.

The above algorithm is inspired by the ACE algorithm. When $\rho(z) = z^2$, the best linear smoother is used at each step. Our algorithm has a broad variety of applications. These include, for example, estimating conditional quantile functions and enhancing robustness.

The sampling behavior of the above method remains unknown. We expect that if the algorithm would be applied to the population instead of to the sample, it would converge to the additive functions. Breiman and Friedman (1985) show that such a statement holds for the case $\rho(z) = z^2$. For a discussion on determining the number of terms in an additive model, see Härdle and Tsybakov (1990).

4.2 Two-term Interaction Nonparametric Regression

The extension of the above method to two-term interaction nonparametric regression is, in some sense, straightforward.

Find the main-effect functions $\phi_j(X_j)$ with

$$E\phi_j(X_j) = 0, \quad j = 1, \cdots, p, \tag{4.4}$$

and the interaction terms $\theta_{i,j}(X_i, X_j)$ with

$$E[\theta_{i,j}(X_i, X_j)|X_i] = 0, \quad \text{and} \quad E[\theta_{i,j}(X_i, X_j)|X_j] = 0, \quad i, j = 1, \cdots, p, \tag{4.5}$$

such that the error not explained by the main effects and the interactions

$$E\rho\left(Y_j - c - \sum_{1}^{p}\phi_j(X_j) - \sum_{i<j}\theta_{i,j}(X_i, X_j)\right) \tag{4.6}$$

is minimized.

The algorithm for estimating the main effects and the interactions is as follows:

step 0 — step 3 are the same as in the previous algorithm.

step 4. Use the residuals from step 3 to estimate the interaction $\theta_{1,2}(x_1, x_2)$ between X_1 and X_2. To do this we apply a two-dimensional local linear regression method which is a natural extension of (3.2). More precisely, $\hat{\theta}_{1,2}(x_1, x_2) = \hat{a}_0$, where the $\hat{a}_i$'s minimize

$$\sum_{j=1}^{n} \rho\left(y_j - a_0 - \sum_{i=1}^{2} a_i(x_i - x_{ji})\right) K\left(\frac{x_1 - x_{j1}}{h_{n1}}, \frac{x_2 - x_{j2}}{h_{n2}}\right),$$

with respect to a_0, a_1 and a_2. In order to cope with the constraints (4.5) we renormalize the resulting estimate in a similar manner as in step 1.

step 5. Use the residuals from step 4 to estimate the interaction $\theta_{1,3}(x_1, x_3)$ between X_1 and X_3, and repeat the same process until all interactions are estimated.

step 6. Repeat steps 1–5 until the mean of the residuals $\frac{1}{n}\sum_{j=1}^{n} \rho(R_j)$ fails to decrease dramatically.

The above algorithm is an extension of the ACE algorithm. This more general algorithm permits us to study both main effects and interactions, and to take into account robustness.

5 Concluding Remarks

We presented a number of advantages of the local linear smoother. These include the adaptation to both random and fixed designs, and to various designs, for example, those for which $|f_X'(x)/f_X(x)|$ is large. Moreover, we proved that with a suitable choice of the kernel and the bandwidth, the local linear smoother is the best linear smoother. For other choices of kernels, computations (see Fan (1991)) show that the resulting local linear smoothers have high efficiency.

Inspired by these optimalities, we extended the above method for more general setups. This lead to studying robust nonparametric regression and to estimating the conditional quantile function—an important quantity for prediction intervals.

To deal with high-dimensional data, the ideas of the best additive approximation and the best two-term-interaction approximation were used. These methods allowed us to examine

both main effects and interactions between covariates. The generalized local linear smoother serves as a building block in these methods.

A number of questions remain open. We mention some of them:

- the behavior of the cross-validation bandwidth (2.4);
- generalized local linear smoothers with variable bandwidth;
- the behavior of the cross-validation bandwidth for the generalized local linear smoother;
- sampling properties of the *iterated robustness enhanced nonparametric estimate*;
- test whether a particular main effect and a interaction exist or not;

A discussion on these issues is beyond the scope of this paper, but introduces further research topics.

6 Proofs

The following two lemmas will be used in the proof of Theorem 1.

Lemma 1. *Let $L(\cdot)$ be a bounded, and uniformly Lipschitz continuous function. Assume that $m(\cdot)$ has a bounded second derivative, and that $G(\cdot)$ has a first derivative which is bounded away from zero, uniformly in its bounded support. Suppose that there exists a $\delta > 0$ such that $nh_n^{1+3\delta} \to \infty$, and $\lim_{|x|\to\infty} x^3 L(x^\delta) = 0$. Then*

$$\frac{1}{n}\sum_{j=1}^{n} L\left(\frac{x - x_j}{h_n}\right) R(x_j) - \int_0^1 L\left(\frac{x - G^{-1}(y)}{h_n}\right) R(G^{-1}(y))dy = o(h_n^3),$$

where $x_j = G^{-1}(j/n)$ and

$$R(x_j) = m(x_j) - m(x) + m'(x)(x - x_j). \tag{6.1}$$

Proof. By the mean-value theorem for integration, we have

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{j=1}^{n} L\left(\frac{x-x_j}{h_n}\right) R(x_j) - \int_0^1 L\left(\frac{x-G^{-1}(y)}{h_n}\right) R(G^{-1}(y))dy\right| \\
\leq\ & \frac{1}{n}\sum_{j=1}^{n}\left|L\left(\frac{x-G^{-1}(j/n)}{h_n}\right) R(G^{-1}(j/n)) - L\left(\frac{x-G^{-1}(\xi_j)}{h_n}\right) R(G^{-1}(\xi_j))\right| \\
\equiv\ & \frac{1}{n}\sum_{j=1}^{n} c_j, \qquad (6.2)
\end{aligned}$$

where $(j-1)/n \leq \xi_j \leq j/n$. Let

$$I = \left\{j : \left|\frac{x-G^{-1}(j/n)}{h_n}\right| \leq a_n\right\},$$

where $a_n = h_n^{-\delta}$. It follows from the assumption on G that

$$\#(I) \leq O(nh_n a_n). \qquad (6.3)$$

Using the mean-value theorem and the conditions on $L(\cdot)$ we have, for all $j \in I$,

$$\begin{aligned}
c_j \leq\ & \left|L\left(\frac{x-G^{-1}(j/n)}{h_n}\right) - L\left(\frac{x-G^{-1}(\xi_j)}{h_n}\right)\right| \left|R(G^{-1}(j/n))\right| \\
& + \left|L\left(\frac{x-G^{-1}(\xi_j)}{h_n}\right)\right| \left|R(G^{-1}(j/n)) - R(G^{-1}(\xi_j))\right| \\
\leq\ & \left(\frac{d_1}{h_n}\left|R(G^{-1}(j/n))\right| + d_2\left|R'(G^{-1}(\eta_j))\right|\right) \left|G^{-1}(j/n) - G^{-1}(\xi_j)\right|,
\end{aligned}$$

where $\xi_j \leq \eta_j \leq j/n$, and d_1, d_2 are some positive constants. Further, by the assumptions on the functions $m(\cdot)$ and $G(\cdot)$, it is easy to see that for all $j \in I$

$$\begin{aligned}
R(G^{-1}(j/n)) &= m(G^{-1}(j/n)) - m(x) + m'(x)(x - G^{-1}(j/n)) \\
&= O\left((x-G^{-1}(j/n))^2\right) \\
&= O\left(a_n^2 h_n^2\right),
\end{aligned}$$

$$R'(G^{-1}(\eta_j)) = m'(G^{-1}(\eta_j)) - m'(x) = O(a_n h_n),$$

and

$$G^{-1}(j/n) - G^{-1}(\xi_j) = O(1/n),$$

where these bounds hold uniformly in $j \in I$. Hence, from (6.3) we obtain that

$$\frac{1}{n}\sum_{j\in I} c_j = O\left(\frac{a_n^3 h_n^2}{n}\right) = o(h_n^3),$$

under the condition on the bandwidth. Finally, by the boundedness of $R(\cdot)$ and the tail condition on $L(\cdot)$, we get

$$\frac{1}{n}\sum_{j\notin I} c_j = O(L(a_n)) = o(h_n^3).$$

This completes the proof. □

Lemma 2. *Let $L(\cdot)$ and $S(\cdot)$ be bounded, and uniformly Lipschitz continuous functions. Assume that $G(\cdot)$ has a first derivative which is bounded away from zero uniformly in its bounded support. Suppose that there exists a $\delta > 0$ such that $nh_n^{1+\delta} \to \infty$, and $\lim_{|x|\to\infty} xL(x^\delta) = 0$. Then*

$$\frac{1}{n}\sum_{j=1}^{n} L\left(\frac{x - x_j}{h_n}\right) S(x_j) - \int_0^1 L\left(\frac{x - G^{-1}(y)}{h_n}\right) S(G^{-1}(y))dy = o(h_n).$$

Proof. The proof follows the same virtue as that of Lemma 1, and is omitted. □

Proof of Theorem 1. First of all, note that the assumption on the tail of K implies that $\int_{-\infty}^{\infty} v^4 K(v)dv < \infty$. By a mean-variance decomposition we find that the MSE of the estimator (2.1) equals

$$\frac{[\sum_{j=1}^n w_j(m(x_j) - m(x))]^2}{(\sum_{j=1}^n w_j)^2} + \frac{\sum_{j=1}^n w_j^2 \sigma^2(x_j)}{(\sum_{j=1}^n w_j)^2}. \tag{6.4}$$

Using Lemma 2 with $L(z) = z^l K(z)$ $(l = 0, 1, 2)$ and $S(\cdot) \equiv 1$, it is easy to show that

$$\begin{aligned}
\frac{s_{n,l}}{nh_n^{l+1}} &= \frac{1}{h_n}\int_0^1 K\left(\frac{x - G^{-1}(y)}{h_n}\right)\left(\frac{x - G^{-1}(y)}{h_n}\right)^l dy + o(1) \\
&= \int_{(x-b)/h_n}^{(x-a)/h_n} K(z) z^l f_X(x - h_n z)dz + o(1) \\
&= s_l f_X(x)(1 + o(1)),
\end{aligned} \tag{6.5}$$

where $s_{n,l}$ is defined in (2.3) and $s_l = \int_{-\infty}^{\infty} v^l K(v)dv$. Thus, we obtain

$$\sum_{j=1}^{n} w_j = s_{n,0}s_{n,2} - s_{n,1}^2 = n^2 h_n^4 s_2 f_X^2(x)(1 + o(1)). \tag{6.6}$$

Application of Lemma 1, with $L = K$, yields

$$\begin{aligned} & \frac{1}{n}\sum_{j=1}^{n} K\left(\frac{x - x_j}{h_n}\right) R(x_j) \\ = & \int_0^1 K\left(\frac{x - G^{-1}(y)}{h_n}\right) R(G^{-1}(y))dy + o(h_n^3), \\ = & \frac{1}{2}m''(x)f_X(x)s_2 h_n^3(1 + o(1)). \end{aligned} \tag{6.7}$$

Again using Lemma 1, with $L(z) = zK(z)$, we show that

$$\begin{aligned} & \frac{1}{n}\sum_{j=1}^{n} K\left(\frac{x - x_j}{h_n}\right)\left(\frac{x - x_j}{h_n}\right) R(x_j) \\ = & \int_0^1 K\left(\frac{x - G^{-1}(y)}{h_n}\right)\left(\frac{x - G^{-1}(y)}{h_n}\right) R(G^{-1}(y))dy + o(h_n^3) \\ = & O(h_n^3). \end{aligned} \tag{6.8}$$

Noting that $\sum_{j=1}^{n} w_j(x - x_j) = 0$ and combining (6.6)—(6.8), it follows that

$$\begin{aligned} \sum_{j=1}^{n} w_j(m(x_j) - m(x)) &= \sum_{j=1}^{n} w_j R(x_j) \\ &= s_{n,2}\sum_{j=1}^{n} K\left(\frac{x - x_j}{h_n}\right) R(x_j) - s_{n,1}\sum_{j=1}^{n} K\left(\frac{x - x_j}{h_n}\right)(x - x_j)R(x_j) \\ &= \frac{1}{2}m''(x)f_X^2(x)s_2^2 n^2 h_n^6(1 + o(1)) \end{aligned} \tag{6.9}$$

Next, we write

$$\begin{aligned} \sum_{j=1}^{n} w_j^2\sigma^2(x_j) &= s_{n,2}^2\sum_{j=1}^{n} K^2\left(\frac{x - x_j}{h_n}\right)\sigma^2(x_j) \\ &\quad -2s_{n,1}s_{n,2}\sum_{j=1}^{n} K^2\left(\frac{x - x_j}{h_n}\right)(x - x_j)\sigma^2(x_j) \\ &\quad +s_{n,1}^2\sum_{j=1}^{n} K^2\left(\frac{x - x_j}{h_n}\right)(x - x_j)^2\sigma^2(x_j). \end{aligned}$$

Now, application of Lemma 2 to each of these three terms leads to

$$\sum_{j=1}^{n} w_j^2\sigma^2(x_j) = n^3 h_n^7 \sigma^2(x) f_X^3(x) s_2^2 \int_{-\infty}^{\infty} K^2(v)dv(1 + o(1)). \tag{6.10}$$

The result follows from (6.4), (6.6), (6.9) and (6.10). □

Proof of Theorem 3. Starting from (2.5), it can be shown that, for each $f \in \mathcal{C}_2$, the MSE of the estimator $\hat{m}^*(x)$ is less than or equal to r. Thus, it remains to establish the lower bound.

Assume without loss of generality that the supremum $\sup_{f\in\mathcal{C}_2} \sigma_m^2(x)$ is attained at a constant regression function $m_0(\cdot) = \theta_0$. Otherwise, one can find a constant regression function $m_0(\cdot)$ such that $\sigma_{m_0}^2(x) \geq a - \varepsilon$, for any $\varepsilon > 0$, and then let $\varepsilon \to 0$ at the last step of the proof.

Let $m_1(z) = \theta_0 - g_n(z)$, $m_2(z) = \theta_0 + g_n(z)$, where g_n satisfies

$$|g_n(y) - g_n(x) - g_n'(x)(y-x)| \leq C(y-x)^2.$$

Denote the conditional MSE of the linear smoother $\hat{m}_L$ (defined in (2.10)) by

$$R(m, \hat{m}_L) = \left[\sum_{j=1}^n W_j m(X_j) - m(x)\right]^2 + \sum_{j=1}^n W_j^2 \sigma_m^2(X_j).$$

Then,

$$\begin{aligned}
&\sup_{f\in\mathcal{C}_2} R(m, \hat{m}_L) \\
\geq\ & \frac{1}{2}[R(m_1, \hat{m}_L) + R(m_2, \hat{m}_L)] \\
=\ & \frac{1}{2}\left[[\sum_{j=1}^n W_j m_1(X_j) - m_1(x)]^2 + [\sum_{j=1}^n W_j m_2(X_j) - m_2(x)]^2\right] \\
& + \frac{1}{2}\sum_{j=1}^n W_j^2 \left[\sigma_{m_1}^2(X_j) + \sigma_{m_2}^2(X_j)\right] \\
\geq\ & \left[\sum_{j=1}^n W_j g_n(X_j) - g_n(x)\right]^2 + \frac{1}{2}\sum_{j=1}^n W_j^2 \left[\sigma_{m_1}^2(X_j) + \sigma_{m_2}^2(X_j)\right], \qquad (6.11)
\end{aligned}$$

where the last inequality follows from the fact that

$$a^2 + b^2 \geq \frac{1}{2}(a-b)^2.$$

Denote $s(X_j) = [\sigma^2_{m_1}(X_j) + \sigma^2_{m_2}(X_j)]/2$. Applying Lemma 1 of Fan (1991), we find that (6.11) is greater than or equal to

$$\frac{g_n^2(x)}{1 + \sum_{j=1}^n g_n^2(X_j)/s(X_j)} = \frac{g_n^2(x)}{1 + nEg_n^2(X_1)/s(X_1) + O_p(\sqrt{nEg_n^4(X_1)/s^2(X_1)})}. \qquad (6.12)$$

Take

$$g_n(y) = \frac{b_n^2}{2}\left[1 - 2C(y-x)^2/b_n^2\right]_+, \quad \text{with} \quad b_n = \left(\frac{15\sqrt{2C}a}{nf_X(x)}\right)^{1/5}.$$

Then, by the continuity assumption on σ_m, we have

$$\lim_{n\to\infty} \sigma_{m_1}(x) = \sigma_{m_0}(x) = a, \quad \lim_{n\to\infty} \sigma_{m_2}(x) = a.$$

The rest of the proof consists of computing (6.12). For details concerning a similar computation, see the proof of Theorem 3 in Fan (1991). □

References

[1] Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Amer. Statist. Assoc.*, **80**, 580-619.

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

[3] Chu, C.K. and Marron, J.S. (1990). Choosing a kernel regression estimator. *Manuscript*

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

[5] Donoho, D. L. and Johnstone, I. M. (1989). Projection-based approximation and a duality with kernel methods. *Ann. Statist.*, **17**, 58-106.

[6] Donoho, D. L. and Liu, R. C. (1991a). Geometrizing rate of convergence I. *Ann. Statist.*, **19**, to appear.

[7] Donoho, D. L. and Liu, R. C. (1991b). Geometrizing rate of convergence II. *Ann. Statist.*, **19**, to appear.

[8] Fan, J. (1990). A remedy to regression estimation and nonparametric minimax efficiency. *Institute of Statistics Mimeo Series*, #2028, Department of Statistics, University of North Carolina, Chapel Hill.

[9] Fan, J. (1991). Design-adaptive nonparametric regression. *Institute of Statistics Mimeo Series*, #2049, Department of Statistics, University of North Carolina, Chapel Hill.

[10] Fan, J. and Gijbels, I. (1991). Variable bandwidth and local linear regression smoothers. *Institute of Statistics Mimeo Series, #2052*, Department of Statistics, University of North Carolina, Chapel Hill.

[11] Fan, J., Hu, T.C. and Truong, Y.K. (1991). Design-adaptive nonparametric function estimation: a unified approach. *Manuscript.*

[12] Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.*, **76**, 817-823.

[13] Gasser, T. and Müller, H.G. (1979). Kernel estimation of regression functions. In Smoothing techniques for curve estimation. Lectures Notes in Math. 757, 23–68, Springer–Verlag, New York.

[14] Hall, P. (1989). On projection pursuit regression. *Ann. Statist.*, **17**, 573-588.

[15] Hall, P. and Jones, M.C. (1990). Adaptive M-estimation in nonparametric regression. *Ann. Statist.*, **18**, 1712-1728.

[16] Härdle, W. and Gasser, T. (1984). Robust non-parametric function fitting. *J. Roy. Statist. Soc. Ser. B*, **46**, 42-51.

[17] Härdle, W. and Tsybakov, A. B. (1990). How many terms should be added into an additive model? CORE discussion paper #9068, University of Leuven, Louvain-la-Neuve, Belgium.

[18] Huber, C. (1981). *Robust Statistics.* Wiley, New York.

[19] Jennen-Steinmetz, C. and Gasser, T. (1988). A unifying approach to nonparametric regression estimation. *Jour. Amer. Statist. Assoc.*, **83**, 1084-1089.

[20] Lehmann, E.L. (1983) *Theory of Point Estimation.* John Wiley and Sons, New York.

[21] Mack, Y.P. and Müller, H.G. (1989). Convolution type estimators for nonparametric regression estimation. *Statist. Probab. letters*, **7**, 229-239.

[22] Müller, H.G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *Jour. Amer. Statist. Assoc.*, **82**, 231-238.

[23] Nadaraya, E.A. (1964). On estimating regression. *Theory Probab Appli.*, **9**, 141-142.

[24] Stone, C.J. (1977). Consistent Nonparametric Regression. *Ann. Statist.*, **5**, 595-620.

[25] Stone, C.J. (1990). L_2 rate of convergence for interaction spline regression. *Tech Report #268*, Dept. of Statist., Univ. of California, Berkeley.

[26] Tsybakov, A.B. (1986). Robust reconstruction of functions by the local-approximation method. *Problems Information Transmissions*, 133-146.

[27] Watson, G.S. (1964). Smooth regression analysis. *Sankhyā*, Ser. A, **26**, 359-372.

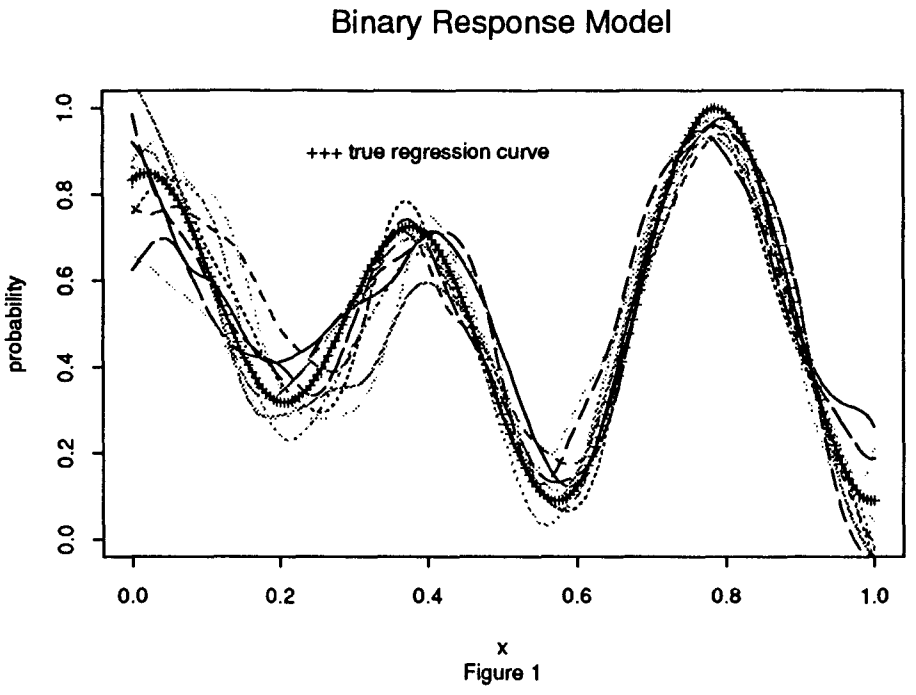

Figure 1

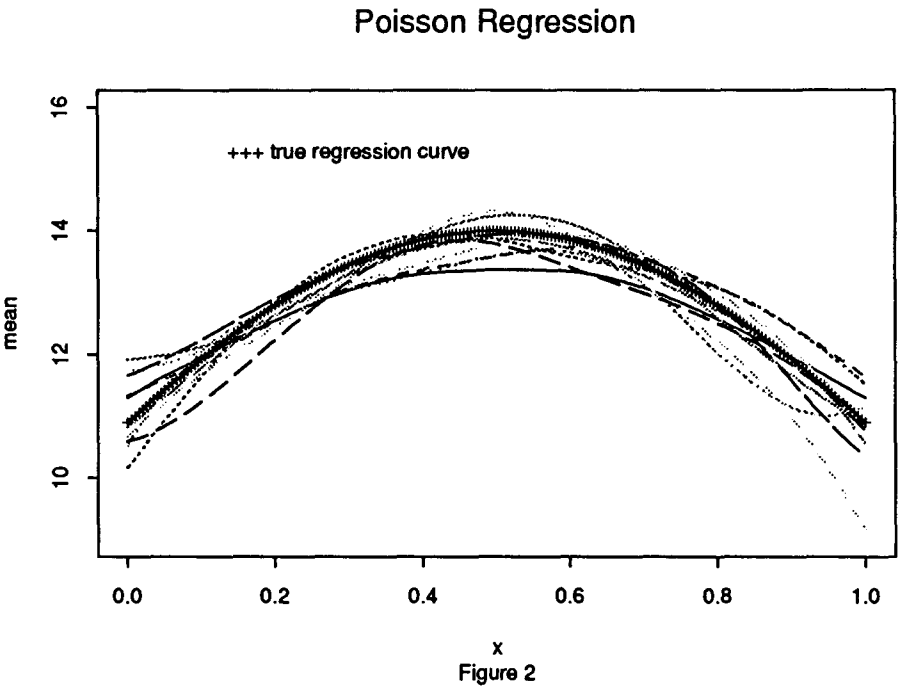

Figure 2

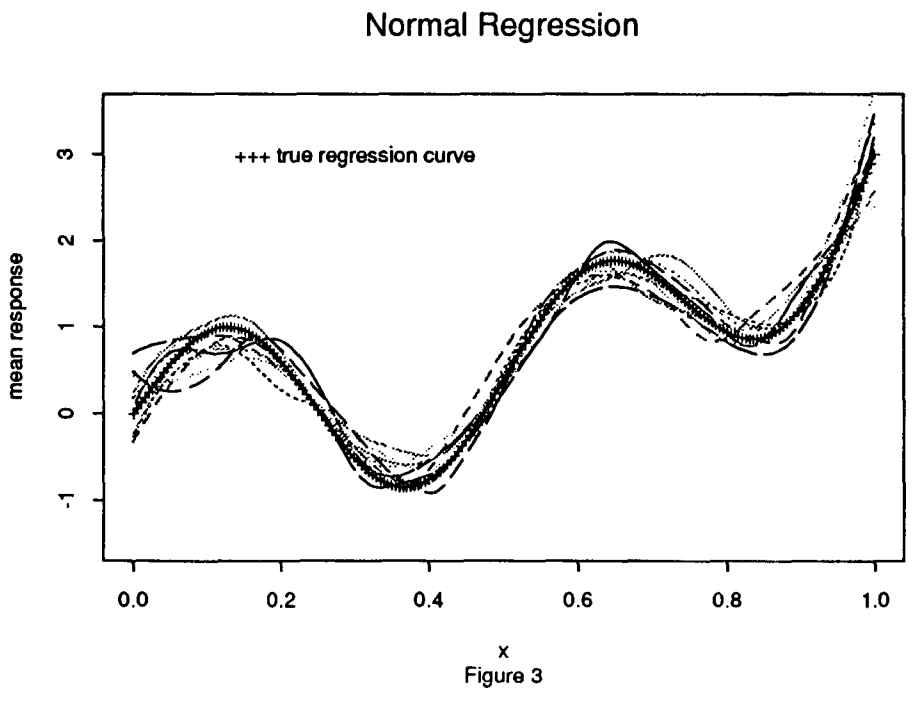

Figure 3