THE INSTITUTE OF STATISTICS

THE CONSOLIDATED UNIVERSITY OF NORTH CAROLINA

LOCAL POLYNOMIAL FITTING: A STANDARD FOR NONPARAMETRIC REGRESSION

by

J. Fan T. Gasser

I. Gijbels M. Brockmann J. Engel

June 1993

Mimeo Series #2302

DEPARTMENT OF STATISTICS
Chapel Hill, North Carolina

Local Polynomial Fitting: A Standard for Nonparametric Regression[1]

Jianqing Fan[2]
Department of Statistics
University of North–Carolina
Chapel Hill, N.C. 27599–3260,
USA

Theo Gasser
Biostatistics Dept., ISPM
University of Zürich
CH–8006 Zürich, Switzerland

Irène Gijbels
Institut de Statistique
Université Catholique de Louvain
Voie du Roman Pays, 34
B–1348 Louvain–la–Neuve, Belgium

Michael Brockmann[1] Joachim Engel[1]
Inst. Appl. Math.
University of Heidelberg
Im Neuenheimer Feld 294
D–6900 Heidelberg, Germany

June 16, 1993

Abstract. Among the various nonparametric regression methods, weighted local polynomial fitting is the one which is gaining increasing popularity. This is due to the attractive minimax efficiency of the method and to some further desirable properties such as the automatic incorporation of boundary treatment. In this paper previous results are extended in two directions: in the one–dimensional case, not only local linear fitting is considered but also polynomials of other orders and estimating derivatives. In addition to deriving minimax properties, optimal weighting schemes are derived and the solution obtained at the boundary is discussed in some detail. An equivalent kernel formulation serves as a tool to derive many of these properties. In the higher dimensional case local linear fitting is considered. Properties in terms of minimax efficiency are derived and optimal weighting

[1]Supported by the Deutsche Forschungsgemeinschaft.
[2]Supported by NSF grant DMS–9203135
We would like to thank Burkhardt Seifert for intellectual support, in particular in the improvement of section 3.

functions both in terms of support and shape are obtained. Asymptotic minimax efficiency is in both cases 100% among the linear estimators and only a small loss has to be tolerated beyond this class.

Key words. Curve estimation, local polynomials, minimax efficiency, minimum least squares, multivariate curve estimation, nonparametric regression

Abbreviated title. Local polynomial regression.

AMS 1991 subject classification. Primary 62G07. Secondary 62C20, 62J20.

1 Introduction

Nonparametric regression techniques have gained increasing popularity, as is documented by the monographs of Eubank (1988), Müller (1988), Härdle (1990) and Wahba (1990). Basically, the form of the regression is dictated by the model when applying parametric regression, and by the data for nonparametric regression. Even when parametric modeling is the ultimate goal, nonparametric methods can prove useful for an exploratory analysis and for checking and improving functional models.

For response variables $Y_1, \ldots, Y_n \in \mathbb{R}$, there are explanatory variables $X_1, \ldots, X_n \in \mathbb{R}^d$, $(X_1, Y_1), \ldots, (X_n, Y_n)$ being independent and identically distributed random variables for the random design model with regression function m

$$m(x) = E(Y|X = x). \tag{1}$$

Non–random $X_1, \ldots, X_n$ are assumed for the following fixed design model:

$$Y_i = m(X_i) + \varepsilon_i \tag{2}$$

where the ε_i are independent and satisfy $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2(X_i)$. While the X come from a d–dimensional probability density with support $\mathrm{Supp}(f_X)$ in model (1), a "regular design" is assumed for (2), leading formally also to a design density f_X on some domain D.

Of interest is the estimation of the regression function m and its derivatives. Most of the literature deals with the one–dimensional case $d = 1$. Popular linear estimators are smoothing splines (Wahba, 1990), kernel estimators of the evaluation type (Nadaraya, 1964; Watson, 1964) and of the convolution type (Gasser and Müller, 1984) and local polynomial fitting (for references see Müller, 1988). The merits of the various methods have been controversial. Splines are close to kernel estimators (Silverman, 1984) with an advantage in terms of minimax properties for the latter (Jennen–Steinmetz and Gasser, 1988). The relative merits of evaluation and convolution weights for kernel estimation have been discussed by Chu and Marron (1991). In summary, the evaluation weights lead to an undesirable form of the bias (Gasser and Engel, 1990; Fan, 1993), while convolution weights pay a price in variance for random designs. In view of the vast literature on the subject it seems timely to offer to the scientific community a method which is attractive on theoretical and on practical grounds and this paper is a contribution in this spirit. Weighted local polynomials are superior to previous methods in efficiency and have some further advantages:

- Local linear fits (for $d = 1$) achieve full minimax efficiency among linear estimators (Fan, 1993) (for the convolution kernel estimators this holds for fixed design only).

- Minimax efficiency remains at 89.4% among all estimators (Fan, 1993).

- The bias at the boundary stays automatically of the same order as in the interior, without the use of specific boundary kernels (Lejeune, 1985; Fan and Gijbels, 1992a).

- The method is appealing on general scientific grounds: 1. The good local approximation properties of polynomials have a deep foundation in mathematics. 2. The least squares principle to be applied opens the way to a wealth of statistical knowledge and thus to easy generalizations. In particular, global polynomial

fitting has always been a popular curve fitting method, allowing it to draw from this pool of knowledge.

This is a big incentive to study the method in some generality. Fast computation was so far possible only in special cases (Friedman, 1984); in a seperate report we present a general and fast algorithm working in $O(n)$ operations (Seifert et al.,1993). Here, we would like to establish a framework for local polynomial fitting, as well as optimality properties. With respect to the first topic, there is some overlap with Müller (1988) and with Ruppert and Wand (1992).

The paper is organized as follows: Section 2 presents the important one–dimensional model, which is formally also easier to understand. Section 3 generalizes the results to higher dimensions for the case of local linear fitting, and Section 4 establishes minimax efficiency.

2 Asymptotic results and optimal weight functions in one dimension

A local polynomial regression at x_0 is computed by minimizing

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} b_j(x_0) \, (X_i - x_0)^j \right)^2 K\left(\frac{X_i - x_0}{h_n} \right) \tag{3}$$

where $K(\cdot)$ denotes a weight function and h_n is a smoothing parameter or bandwidth. Denote by $\widehat{b}_j(x_0) \quad (j = 0, \ldots, p)$ the solution of the least squares problem (3). By Taylor's formula $\widehat{m}_\nu(x_0) = \nu! \widehat{b}_\nu(x_0)$ is an estimator for $m^{(\nu)}(x_0)$. Putting

$$X = \begin{pmatrix} 1 & (X_1 - x_0) & \cdots & (X_1 - x_0)^p \\ \vdots & \vdots & & \vdots \\ 1 & (X_n - x_0) & \cdots & (X_n - x_0)^p \end{pmatrix}, \quad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \widehat{b}(x_0) = \begin{pmatrix} \widehat{b}_0(x_0) \\ \vdots \\ \widehat{b}_p(x_0) \end{pmatrix}$$

and

$$W = \operatorname{diag}\left(K\left(\frac{X_i - x_0}{h_n} \right) \right),$$

the $n \times n$ diagonal matrix of weights, the solution to the least squares problem (3) can be written as

$$\begin{aligned} \widehat{b}(x_0) &= (X^T W X)^{-1} X^T W Y \qquad (4) \\ &= \begin{pmatrix} S_{n,0}(x_0) & S_{n,1}(x_0) & \cdots & S_{n,p}(x_0) \\ S_{n,1}(x_0) & S_{n,2}(x_0) & \cdots & S_{n,p+1}(x_0) \\ \vdots & \vdots & \ddots & \vdots \\ S_{n,p}(x_0) & S_{n,p+1}(x_0) & \cdots & S_{n,2p}(x_0) \end{pmatrix}^{-1} \begin{pmatrix} T_{n,0}(x_0) \\ T_{n,1}(x_0) \\ \vdots \\ T_{n,p}(x_0) \end{pmatrix} \equiv S_n^{-1} T_n \end{aligned}$$

where

$$\begin{aligned} S_{n,j}(x_0) &= \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)^j, \quad j = 0, 1, \ldots, 2p, \\ T_{n,j}(x_0) &= \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)^j Y_i, \quad j = 0, 1, \ldots, p. \end{aligned}$$

Hence

$$\widehat{b}_\nu(x_0) = e_\nu^T \widehat{b}(x_0) = \sum_{i=1}^{n} W_\nu^n\left(\frac{X_i - x_0}{h_n}\right) Y_i \qquad (5)$$

where $e_\nu = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ with 1 at the $(\nu+1)$th position and $W_\nu^n(t) = e_\nu^T S_n^{-1}(1, th_n, \ldots, (th_n)^p)^T K(t)$.

The inverse matrix in (4) exists when the weight function K is non–negative and the sample contains at least p distinct X_i with non–negative weights. This is the case of interest as will be shown below. The following lemma is an important one, as it gives a representation in terms of an equivalent kernel estimator. It proves to be useful for deriving bias and variance in what follows, and minimax properties later on.

Lemma 1 *Let us assume that the design density $f_X(x_0)$ is continuous and positive at x_0. Then*

$$\widehat{b}_\nu(x_0) = \frac{1}{n h_n^{\nu+1} f_X(x_0)} \sum_{i=1}^{n} K_\nu^*\left(\frac{X_i - x_0}{h_n}\right) Y_i \,(1 + o_P(1)) \qquad (6)$$

where

$$K_\nu^*(t) = e_\nu^T S^{-1}(1, t, \ldots, t^p)^T K(t) = \sum_{l=0}^{p} S^{\nu l} t^l K(t) \tag{7}$$

with $S = \left(\int t^{j+l} K(t)dt\right)_{0\le j,l\le p}$ *and* $S^{-1} = \left(S^{jl}\right)_{0\le j,l\le p}$.

The equivalent kernel K_ν^* is a kernel of order $(\nu, p+1)$ up to normalizing constants (Gasser, Müller and Mammitzsch, 1985), which can be shown by checking the definition of these kernels. Equivalent kernels have been previously used for analyzing polynomial fitting by Lejeune (1985) and Müller (1987) in a slightly different way.

This leads directly to bias and variance, and therefore to mean squared error, of the estimator $\widehat{m}_\nu(x_0)$ (see Appendix A.1):

Theorem 1 *Let us assume that* $m^{(p+1)}(x_0)$ *is bounded and that* $(p-\nu)$ *is odd. Bias and variance of* $\widehat{m}_\nu(x_0)$ *are obtained as*

$$\mathrm{E}(\widehat{m}_\nu(x_0)|X_1, \ldots, X_n) - m^{(\nu)}(x_0) = \left(\int t^{p+1} K_\nu^*(t)dt\right) \frac{\nu!\, m^{(p+1)}(x_0)}{(p+1)!} h_n^{p+1-\nu} \left(1+o_P(1)\right)$$

$$\mathrm{Var}(\widehat{m}_\nu(x_0)|X_1, \ldots, X_n) = \frac{\nu!^2 \sigma^2(x_0)}{n h_n^{2\nu+1} f_X(x_0)} \int K_\nu^{*2}(t)dt \ \left(1+o_P(1)\right).$$

Remark 1 *For polynomial fitting p should be of the order* $p = \nu+1, \nu+3, \ldots$ *(Fan and Gijbels, 1992b). Contrary to expectation, the use of a lower order polynomial* $p^* = \nu, \nu+2, \ldots$ *with fewer parameters does not lead to a smaller asymptotic variance. However, an additional lower order bias arises depending on* $m^{(p^*+1)}$ *and also on* f_X *and* f_X'. *A classical example is the Nadaraya–Watson estimator obtained by local constant fitting. Henceforth, we assume* $(p-\nu)$ *odd.*

Remark 2 *The convolution kernel estimator has the same bias, and for fixed design also the same variance, but an unconditional variance which is higher by a factor 1.5 for random design (Gasser and Müller, 1984, Chu and Marron, 1991). Note that the order p of the polynomial correponds to the order k of the kernel* W_k *of the convolution estimator as* $p = k-1$ *(Müller, 1988).*

Remark 3 *The weights in (5) satisfy the following discrete orthogonality relation:*

$$\sum_{i=1}^{n}(X_i - x_0)^q W_\nu^n\left(\frac{X_i - x_0}{h_n}\right) = \delta_{\nu,q} \quad 0 \le \nu, q \le p$$

which leads to zero finite bias for polynomials up to order p. Moment conditions and the respective zero bias are satisfied only asymptotically for convolution kernel estimators.

By using Theorem 1, the asymptotically optimal local bandwidth is obtained as

$$h_{\mathrm{opt}}(x_0) = \left(\frac{(p+1)!^2(2\nu+1)\int K_\nu^{*2}(t)dt\,\sigma^2(x_0)}{2\,n\,(p+1-\nu)(\int t^{p+1}K_\nu^*(t)dt)^2\left(m^{(p+1)}(x_0)\right)^2 f_X(x_0)}\right)^{1/(2p+3)}$$

and the respective global bandwidth as

$$h_{\mathrm{opt}} = \left(\frac{(p+1)!^2(2\nu+1)\int K_\nu^{*2}(t)dt\int \sigma^2(x)/f_X(x)w(x)dx}{2\,n\,(p+1-\nu)(\int t^{p+1}K_\nu^*(t)dt)^2\int\left(m^{(p+1)}(x)\right)^2 w(x)dx}\right)^{1/(2p+3)}$$

for some weight function $w \ge 0$. It is understood that the denominators do not vanish.

A point at the boundary of the form $x = a + ch_n$ $(c \ge 0)$ can be treated similarily. The moments are now defined by $s_{j,c} = \int_{-c}^{\infty} u^j K(u)du$, leading to an equivalent boundary kernel (compare with (7))

$$K_{\nu,c}^*(t) = e_\nu^T S_c^{-1}(1, t, \ldots, t^p)^T K(t) \text{ with } S_c = (s_{j+l,c})_{0\le j,l\le p}.$$

Thus the polynomial method adapts automatically for boundary effects, whereas specific boundary kernels had to be derived for the kernel method (Gasser, Müller and Mammitzsch, 1985).

Next, the question arises which weight function should be used for different choices of ν and p. The asymptotically mean squared error (MSE) depends on the weight function through

$$T_\nu(K) \equiv \left|\int t^{p+1}K_\nu^*(t)dt\right|^{2\nu+1}\left(\int K_\nu^{*2}(t)dt\right)^{p+1-\nu}. \tag{8}$$

"Optimal" kernels K_ν^*, minimizing the right hand side of (8), have been derived by Gasser, Müller and Mammitzsch (1985) and Granovsky and Müller (1991), postulating a minimal number of sign changes to avoid degeneracy. The following theorem provides a simple solution in the context of polynomial fitting:

Theorem 2 *The Epanechnikov weight function* $K(x) = 3/4(1-x^2)_+$ *is the optimal kernel in the sense that it minimizes* $T_\nu(K)$ *over all non-negative functions. It also induces kernels* K_ν^* *which are optimal in the sense of Gasser, Müller and Mammitzsch (1985). Similarly, minimum variance kernels minimizing* $\int K_\nu^{*2}(t)dt$ *are induced by uniform weights* $1/2\mathrm{I}_{\{|x|\leq 1\}}$.

The proof is given in Appendix A.2 and follows also from the minimax results in Section 4.

The question then arises whether these properties carry over to the boundary when using the Epanechnikov or uniform weight function for local polynomial fitting. The appropriate kernels to be used at the boundary are kernels with support $[-1, c]$ satisfying the appropriate moment conditions ("boundary kernels"). In this way, the same bias rates hold for the boundary as for the interior. When using the uniform weight function for local polynomial fitting, the induced boundary kernels achieve still the minimum variance property. The behavior is thus optimal in this sense for the interior as well as for the boundary (Gasser, Müller and Mammitzsch, 1985). This follows immediately from a characterization of minimum variance kernels given by Müller (1991).

The situation is not as simple when taking mean squared error instead of variance as the criterion of optimality. In the context of kernel estimation there is so far no convincing solution for optimal boundary kernels available. This makes it also difficult to judge the quality of the Epanechnikov weight function, when fitting local polynomials at the boundary. Side conditions are necessary to make the variational problem for kernels well–posed. Two approaches have been proposed and used:

(1) Increase the order of the polynomial kernel by one at the boundary compared to the optimal interior kernel and make the kernel vanish at -1 and c (Müller, 1991).

Note that the zero at c is rather unintuitive and has been mainly introduced to use the theory for minimum variance kernels.

(2) Leave the order of the polynomial kernel at the boundary as it is in the interior and put it to zero at -1.

Here, the condition of an invariant order is not convincing given the neat results for minimum variance kernels, where the order at the boundary is automatically increased by one. The equivalent boundary kernels when fitting local polynomials with the Epanechnikov weight function are characterized as follows:

(3) Increase the order of the equivalent polynomial kernel by one, compared to the interior, and put it to zero at -1 and 1.

For the second order boundary kernels ($k = 2$, $\nu = 0$) we then obtain

$$\begin{aligned} K_1(c,x) &= \frac{6(1+x)(c-x)}{(1+c)^3}\left[1+\left(\frac{1-c}{1+c}\right)^2+10\frac{1-c}{(1+c)^2}x\right] \\ K_2(c,x) &= \frac{6(1+x)}{(1+c)^4}\left[1-2c+3c^2+(2-4c)x\right] \\ K_3(c,x) &= \frac{12(1-x^2)}{(1+c)^4(19-18c+3c^2)}\left[8-16c+24c^2-12c^3+(15-30c+15c^2)x\right] . \end{aligned}$$

These kernels are depicted in Figures 1–3. Note the striking qualitative similarity between K_2 and K_3. To compare their

Put Figures 1–3, Table 1 around here

performance we computed the asymptocically optimal mean squared error, integrated over the boundary while assuming $m^{(k)}$ to be constant over the boundary. The

first mentioned solution performs badly, whereas the other two are very close to each other. Thus the induced boundary kernels of the polynomial method are also well-behaving with respect to integrated mean squared error even if optimality cannot be guaranteed.

3 Local linear fitting in higher dimensions

The usefulness of local polynomial fitting in higher dimensions is illustrated for the local linear case ($p = 1$, $\nu = 0$). This is the case of most practical interest, since the sparsity of data in higher dimensions becomes more of a problem for higher order polynomials ("curse of dimensionality"). In principle, the methodology generalizes to higher order polynomials using a multi–indices notation. However, some asymptotic properties regarding optimal bandwidths and optimal weight functions do not necessarily carry over. In this section, bias, variance and optimal weighting functions are derived, whereas minimax properties are discussed in Section 4.

Let K be a function from $\mathbb{R}^d$ in $\mathbb{R}$ and denote by $x = (x_1, \ldots, x_d)$ a point in the d–dimensional space. The observations are given by $(X_1, Y_1), \ldots, (X_n, Y_n)$, with $X_j = (X_{j1}, \ldots, X_{jd})$, $j = 1, \ldots, n$. Consider

$$X = \begin{pmatrix} 1 & X_{11} - x_1 & \cdots & X_{1d} - x_d \\ 1 & X_{21} - x_1 & \cdots & X_{2d} - x_d \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} - x_1 & \cdots & X_{nd} - x_d \end{pmatrix} \quad \widehat{b}(x) = \begin{pmatrix} \widehat{b}_0(x) \\ \widehat{b}_1(x) \end{pmatrix}$$

and denote

$$K_B(u) = \frac{1}{|B|} K(B^{-1}u),$$

where B is a nonsingular matrix, called bandwidth matrix. Further let

$$W = \operatorname{diag}\left(K_B(X_1 - x), \ldots, K_B(X_n - x)\right)$$

be the diagonal matrix of weights.

The local linear regression estimator $\widehat{m}(x) = e_1^T \widehat{b}(x)$ is obtained via (4). Asymptotic analysis yields the following expression for the conditional mean squared error:

$$\begin{aligned} \mathrm{E}\left((\widehat{m}(x) - m(x))^2 \,|X_1, \ldots, X_n\right) &= \left[\frac{1}{4}\left(\mathrm{tr}\left\{H(x)BB^T \int K(u)uu^T du\right\}\right)^2 \right. \\ &\quad \left. + \frac{1}{n|B|}\int K^2(u)du \frac{\sigma^2(x)}{f_X(x)}\right](1 + o_P(1)), \end{aligned}$$

where $H(x)$ stands for the Hessian matrix of m at x (cf. Ruppert and Wand, 1992).

Without loss of generality K satisfies

$$\int u_i u_j K(u)du = \delta_{ij}\mu_2(K)$$

where $\mu_2(K)$ is a positive constant. Differentiating the conditional mean squared error with respect to the matrix BB^T leads to the necessary condition for a local minimum

$$\frac{1}{2}\mu_2^2(K)\mathrm{tr}(HBB^T)H - \frac{M(K)\sigma^2(x)}{2nf_X(x)|B|}(BB^T)^{-1} = 0$$

(see Rao, 1973, p. 72), where $M(K) = \int K^2(u)du$. If the Hessian $H(x)$ is positive or negative definite this equation has a unique solution

$$BB^T = \left(\frac{M(K)\sigma^2(x)|H^*|^{1/2}}{\mu_2^2(K)ndf_X(x)}\right)^{\frac{2}{d+4}}(H^*)^{-1} \tag{9}$$

with

$$H^* = \begin{cases} H & \text{for positive definite } H \\ -H & \text{for negative definite } H \end{cases}$$

which constitutes a minimum. The bandwidth matrix B itself can be chosen as any matrix satisfying the last equation. As can be seen from (10), the MSE does not depend on the particular choice of B. Moreover, if the weight function is rotation invariant (see below), then even the estimate does not depend on the choice of B. Relation (9) leads to insight into the problem of multivariate bandwidth choice: when performing an eigenvalue decomposition of H^*, one gets first an optimal rotation of

the coordinate system, aligning according to the Hessian matrix. The respective eigenvalues lead to a scaling of the weight function K in direction of the new axes, similar to the analogous problem of bandwidth choice in one dimension. The case of an indefinite H comprises features which are characteristic for dimensions higher than one. It is then possible to choose the bandwidths in different directions appropriately such that the above leading term of the bias vanishes. In such regions with lower order bias, it would be necessary to perform a higher order analysis. In addition, zero eigenvalues of the Hessian matrix lead to problems similar to the linear case in one dimension. A detailed discussion is beyond the scope of this paper.

Substituting the optimal choice for the bandwidth matrix into the asymptotic expression for the MSE, we get

$$\mathrm{AMSE} = \frac{d+4}{4} d^{\frac{4}{d+4}} \left(M^2(K)\mu_2^d(K)\right)^{\frac{2}{d+4}} \left(\frac{\sigma^4(x)}{n^2 f_X^2(x)} |H^*(x)|\right)^{\frac{2}{d+4}}. \tag{10}$$

The above formula leads to the formalization of optimal weight functions: Find K such that

$$M^2(K)\mu_2^d(K) = \left(\int K^2(u)du\right)^2 \int u_1^2 K(u)du \cdots \int u_d^2 K(u)du \tag{11}$$

is minimized subject to

$$\int K(u)du = 1,\ \int uK(u)du = 0,\ K \geq 0,\ \int u_i u_j K(u)du = \delta_{ij}\mu_2(K).$$

The solution given in the next theorem is derived in Appendix A.3.

Theorem 3 *The optimal weight function is the spherical Epanechnikov kernel*

$$K_0(u) = \frac{d(d+2)}{2S_d}(1 - u_1^2 - \cdots - u_d^2)_+,$$

where $S_d = 2\pi^{d/2}/\Gamma(d/2)$ denotes the area of the surface of the d-dimensional unit ball.

Thus the theorem also provides an answer to the open problem about the optimal support of the weight function in higher dimensions. With the optimal weight function, we can easily obtain that

$$\mu_2(K_0) = \frac{1}{d+4} \quad \text{and} \quad M(K_0) = \frac{2d(d+2)}{(d+4)S_d}.$$

These two moments are useful in bandwidth selection and for risk calculation.

4 Efficiency considerations

In this section minimax efficiency of local polynomial estimators is studied, generalizing results obtained by Fan (1993) for the special case $d = 1$, $p = 1$, $\nu = 0$. First, it is shown that these estimators achieve full efficiency among all linear estimators and second that a relatively small loss in efficiency occurs when allowing nonlinear estimators as well. For an illuminating account on recent developments of minimax theory, see Donoho (1990), Donoho and Liu (1991) and Donoho and Johnstone (1992) where attention is focussed on density estimation and white noise models.

4.1 Estimating $m^{(\nu)}$ in one dimension

Most popular function estimators are linear admitting the representation $\sum_{i=1}^{n} w_i(X_1, \ldots, X_n)Y_i$. In this section, previous minimax results for local linear fitting (Fan, 1992, 1993) are extended to higher order polynomials and to derivative estimation.

Without loss of generality, our goal is to estimate the functional $S_\nu(m) = m^{(\nu)}(0)$, where 0 represents an arbitrary interior point. Let

$$\mathcal{C} = \{m : |m(z) - \sum_{j=0}^{p} \frac{m^{(j)}(0)}{j!} z^j| \leq C \frac{|z|^{p+1}}{(p+1)!}\} \tag{12}$$

be the class of regression functions whose $(p+1)^{th}$ -derivative is bounded by C.

Further basic conditions are as follows:

Condition A

(a) $\sigma(\cdot)$ is continuous at the point 0,

(b) $f_X(\cdot)$ is continuous at the point 0 with $f_X(0) > 0$,

(c) $p - \nu$ is odd.

Denote by

$$R_L(\nu) = \inf_{\hat{S}_\nu \text{ linear}} \sup_{m \in \mathcal{C}} E\left\{\left(\hat{S}_\nu - m^{(\nu)}(0)\right)^2 | X_1, \cdots, X_n\right\} \tag{13}$$

the linear minimax risk and by

$$R_N(\nu) = \inf_{\hat{T}_\nu} \sup_{m \in \mathcal{C}} E\left\{\left(\hat{T}_\nu - m^{(\nu)}(0)\right)^2 | X_1, \cdots, X_n\right\} \tag{14}$$

the minimax risk.

As has been shown by Donoho and Liu (1991) and Fan (1993) the modulus of continuity

$$\omega_\nu(\varepsilon) = \sup\{|m_1^{(\nu)}(0) - m_0^{(\nu)}(0)| : m_0, m_1 \in \mathcal{C}, \|m_1 - m_0\| = \varepsilon\} \tag{15}$$

is a quantity related to the minimax risk.

Denote the optimal kernel of order $(\nu, p+1)$ by $K^{opt}_{\nu,p+1}(x)$, to be given in Appendix A.4.2, (Gasser, Müller and Mammitzsch, 1985). Note that as shown in Theorem 3, $K^{opt}_{\nu,p+1}(x) = K^*_{0\nu}(x)$ - the equivalent kernel of the Epanechnikov weight function. The L_2-norm of this kernel is given by (A.10). The following two lemmas are essential for proving Theorem 4 below:

Lemma 2 *Let λ_ν be given by (A.15). Then*

$$\omega_\nu(\varepsilon) = 2\nu!|\lambda_\nu|\left(\frac{C}{(p+1)!|\lambda_{p+1}|}\right)^s \|K^{opt}_{\nu,p+1}\|^{-r}\left(\frac{\varepsilon}{2}\right)^r(1 + o(\varepsilon)) \qquad as \quad \varepsilon \to 0,$$

where

$$r = \frac{2(p+1-\nu)}{2p+3}, \qquad s = \frac{2\nu+1}{2p+3}. \tag{16}$$

Lemma 2 together with Theorem 6 of Fan (1993) allow us to evaluate the minimax bounds as follows.

Lemma 3 *The linear minimax risk is given by*

$$R_L(\nu) = r_{\nu,p}(1 + o_P(1)), \tag{17}$$

and the minimax risk is asymptotically bounded by

$$r_{\nu,p} \geq R_N(\nu) \geq (0.894)^2 r_{\nu,p} \tag{18}$$

with

$$\begin{aligned} r_{\nu,p} &= \left(\frac{2p+3}{2\nu+1}\right)\left(\frac{(p+\nu+2)!}{\left(\frac{p+1+\nu}{2}\right)!\left(\frac{p+1-\nu}{2}\right)!}\right)^2\left(\frac{r}{(p+\nu+2)4^{p+2}}\right)^r \\ &\times\left[\begin{pmatrix} 2p+2 \\ p+1 \end{pmatrix}(2p+3)\right]^{-2s}\left(\frac{C}{(p+1)!}\right)^{2s}\left(\frac{\sigma^2(0)}{nf_X(0)}\right)^r. \end{aligned} \tag{19}$$

Let $\hat{m}^{(\nu)}(0) = \hat{b}_\nu(0)$ be the estimator resulting from a local polynomial fit of order p with the Epanechnikov weight function $K_0(t)$ and the optimal bandwidth. This estimator achieves the optimality properties in the previous lemma:

Theorem 4 *The local polynomial fit $\hat{m}^{(\nu)}(0)$ is a best linear estimator for $m^{(\nu)}(0)$ in the sense that*

$$\frac{R_L(\nu)}{\sup_{m\in\mathcal{C}} E\left[\left(\hat{m}^{(\nu)}(0) - m^{(\nu)}(0)\right)^2 | X_1, \cdots, X_n\right]} \xrightarrow{P} 1. \tag{20}$$

Moreover, the estimator has an asymptotic efficiency of at least 89.4 % among all estimators:

$$\frac{R_N(\nu)}{\sup_{m\in\mathcal{C}} E\left[\left(\hat{m}^{(\nu)}(0) - m^{(\nu)}(0)\right)^2 | X_1, \cdots, X_n\right]} \geq (0.894)^2 + o_P(1). \tag{21}$$

The minimax theory is a device to justify the intuitively appealing local polynomial methods. Other justifications such as graphical representation and finite sample simulations can be found in Fan (1992), Hastie and Loader (1993) and Fan and Marron (1993).

4.2 Local linear fitting in higher dimensions

It will be shown that minimax properties carry over to higher dimensions for local linear fitting. To accomplish this, we assume that the regression function is in the class:

$$\mathcal{C}_2 = \{m : |m(z) - (m'(x))(z-x)^T| \leq \frac{1}{2}(z-x)C(z-x)^T\},$$

where C is a positive definite $(d \times d)$-matrix. Intuitively, this class includes regression functions whose Hessian matrix is bounded by C. We use R_L and R_N to denote respectively the minimax linear risk and the minimax risk. These risks are defined similarly to (13) and (14). Without loss of generality, we assume that $x = 0$. Applying Theorem 6 of Fan (1993), we obtain the following results to be proved in Appendix A.5:

Theorem 5 *Suppose that Condition A a) and b) hold. Then, the local linear fit with spherically symmetric Epanechnikov weight function is a best linear estimator and has minimax efficiency at least 89.4% in the sense similar to Theorem 4. Moreover, the linear minimax risk is given by*

$$\begin{aligned} R_L &= \frac{d}{4}\Big(\frac{2}{S_d}\Big)^{\frac{4}{d+4}}(d+2)^{\frac{4}{d+4}}(d+4)^{-\frac{d}{d+4}}\Big(\frac{\sigma^4(0)}{n^2 f_X^2(0)}|C|\Big)^{\frac{2}{d+4}}(1+o_P(1)) \\ &\equiv r_d(1+o_P(1))\,, \end{aligned} \tag{22}$$

and the minimax risk is bounded by

$$r_d \geq R_N \geq (0.894)^2 r_d.$$

REFERENCES

Chu, C.K. and Marron, J.S. (1991). Choosing a kernel regression estimator. *Statistical Science*, **6**, 404–433.

Donoho, D.L. (1990). Statistical estimation and optimal recovery. *Tech. Report 214*, Department of Statistics, University of California, Berkeley.

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

Donoho, D.L. and Liu, R.C. (1991). Geometrizing rate of convergence III. *Ann. Statist.*, **19**, 668–701.

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

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

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

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

Fan, J. and Gijbels, I. (1992b). Spatial and design adaptation: Variable order approximation in function estimation. *Institute of Statistics Mimeo Series # 2080*, Univeristy of North Carolina at Chapel Hill.

Fan, J. and Marron, S. (1993). Comments on "Local regression: Automatic Kernel Carpentry" by Hastie and Loader. *Statistical Sciences*, to appear.

Friedman, J.H. (1984). A variable span smoother. *Techn. Report* , Department of Statistics, Stanford University.

Gasser, T. and Engel, J. (1990). The choice of weights in kernel regression estimation. *Biometrika*, **77**, 377–381.

Gasser, T. and Müller, H.–G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. of Statist.*, **11**, 171–185.

Gasser, T. , Müller, H.–G. and Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *J. Royal Statist. Soc. B*, **47**, 238–252.

Granovsky and Müller, H.–G. (1991), Optimizing kernel methods: a unifying variational principle, *International Statistical Review*, **59**, 373–388.

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

Hastie, T. and Loader, C. (1993). Local regression: Automatic Kernel Carpentry. *Statistical Sciences*, to appear.

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

Lejeune (1985). Estimation non-paramétrique par noyaux: régression polynomiale mobile. *Revue de Statist. Appliq.*, **33**, 43–68.

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

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

Müller, H.–G. (1991). Smooth optimum kernel estimators near endpoints. *Biometrika*, **78**, 521–530.

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

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications.* Wiley, New York.

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

Seifert, B., Brockmann, M., Engel, J. and Gasser, T. (1993). Fast algorithms for nonparametric curve estimation. Manuscript.

Silverman, B.W. (1984). Spline smoothing: the equivalent variable kernel method. *Ann. Statist.*, **12**, 898–916.

Wahba, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

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

Appendix — Proofs of the results

A.1 Proof of Theorem 1

¿From (5) and a Taylor expansion one obtains

$$\begin{aligned} E\big(\hat{b}_\nu(x)|X_1,\cdots,X_n\big) &= \sum_{i=1}^{n} W_\nu^n(\frac{X_i-x}{h_n})\, m(X_i) \\ &= \sum_{i=1}^{n} W_\nu^n(\frac{X_i-x}{h_n}) \sum_{l=0}^{p+1} \frac{m^{(l)}(x)}{l!}(X_i-x)^l + o_P(h_n^{p+1}) \\ &= \frac{m^{(\nu)}(x)}{\nu!} + \sum_{i=1}^{n} W_\nu^n(\frac{X_i-x}{h_n})(X_i-x)^{p+1}\frac{m^{(p+1)}(x)}{(p+1)!} + o_P(h_n^{p+1}). \end{aligned}$$

Hence, the conditional bias of $\hat{b}_\nu(x)$ is given by the second term in this expression and equals

$$e_\nu^T {S_n}^{-1} \begin{pmatrix} S_{n,p+1}(x) \\ \vdots \\ S_{n,2p+1}(x) \end{pmatrix} \frac{m^{(p+1)}(x)}{(p+1)!} \approx e_\nu^T\, HS^{-1}\, H \begin{pmatrix} h_n^{p+1} s_{p+1} \\ \vdots \\ h_n^{2p+1} s_{2p+1} \end{pmatrix} \frac{m^{(p+1)}(x)}{(p+1)!}$$

$$= h_n^{p+1-\nu} e_\nu^T S^{-1} \begin{pmatrix} s_{p+1} \\ \vdots \\ s_{2p+1} \end{pmatrix} \frac{m^{(p+1)}(x)}{(p+1)!}$$

$$= \left(\int t^{p+1} K_\nu^*(t) dt \right) \frac{m^{(p+1)}(x)}{(p+1)!} h_n^{p+1-\nu}, \quad \text{(A.1)}$$

by utilizing the definition of the kernel K_ν^* in (7), and with

$$H = \operatorname{diag}\left(\left(h_n^{-i} \right)_{0 \le i \le p} \right) .$$

For the conditional variance of $\hat{b}_\nu(x)$ we find

$$\operatorname{Var}(\hat{b}_\nu(x) | X_1, \cdots, X_n) \approx e_\nu^T {S_n}^{-1} X^T W^2 X {S_n}^{-1} e_\nu \sigma^2(x) \approx \frac{\sigma^2(x)}{n h_n^{2\nu+1} f_X(x)} e_\nu^T S^{-1} V S^{-1} e_\nu, \quad \text{(A.2)}$$

with $V = (V_{ij})$ and $V_{ij} = \int t^{i+j} K^2(t) dt$, $0 \le i, j \le p$. Further,

$$e_\nu^T S^{-1} V S^{-1} e_\nu = \sum_{i=0}^{p} \sum_{l=0}^{p} S^{i\nu} S^{l\nu} V_{i\nu} = \int K_\nu^{*2}(t) dt. \quad \text{(A.3)}$$

Expressions (A.1) – (A.3) lead to the desired results. □

A.2 Proof of Theorem 2

In order to use the same normalization constant as Gasser, Müller and Mammitzsch (1985), we define the equivalent kernel by (compare with (7))

$$W_\nu^*(x) = (-1)^\nu \nu! e_\nu^T S^{-1} \left(1, x, \ldots, x^p\right)^T K_0(x).$$

The following properties hold for W_ν^*:

1. W_ν^* is a $(\nu, p+1)$–kernel in the sense of Gasser, Müller and Mammitzsch (1985).

2. Obviously, $W_\nu^*(-1) = W_\nu^*(1) = 0$.

3. Because $\nu + p$ is odd, the last element of the ν–th row of S^{-1} is zero and therefore by (7), W_ν^* is a polynomial of degree $p+1$ on $[-1, 1]$.

Theorem C in Granovsky and Müller (1991) states that these properties characterize the optimal kernel. □

A.3 Proof of Theorem 3

Let

$$\varphi(K) = \Big(\int K^2(u)du\Big)^2 \int u_1^2 K(u)du \cdots \int u_d^2 K(u)du\,. \tag{A.4}$$

We had assumed without loss of generality that

$$\int u_1^2 K(u)du = \cdots = \int u_d^2 K(u)du\,, \tag{A.5}$$

which leads to

$$\varphi(K) = \Big(\int K^2(u)du\Big)^2 \Big(\frac{1}{d}\int (u_1^2 + \cdots + u_d^2)K(u)du\Big)^d\,. \tag{A.6}$$

Property (A.6) allows us to find the optimal K through the following minimization problem:

$$\begin{aligned}&\text{Minimize } \textstyle\int K^2(u)du \text{ subject to}\\ &\textstyle\int K(u)du = 1, \quad K \geq 0, \quad \int uK(u)du = 0, \quad \int u_i u_j K(u)du = 0 \quad \text{when } i \neq j\\ &\textstyle\int (u_1^2 + \cdots + u_d^2)K(u)du = \int (u_1^2 + \cdots + u_d^2)K_0(u)du.\end{aligned} \tag{A.7}$$

Now for any nonnegative kernel $K \geq 0$, let $\delta = K - K_0$.

$$\int \|u\|^2 \delta(u)du = 0, \qquad \int \delta(u)du = 0.$$

Hence, we find that

$$\begin{aligned}\int \delta(u)K_0(u)du = \int_{\{\|u\|^2 \leq 1\}} \delta(u)(1 - \|u\|^2)du &= -\int_{\{\|u\|^2 > 1\}} \delta(u)(1 - \|u\|^2)du\\ &= \int_{\{\|u\|^2 > 1\}} K(u)(\|u\|^2 - 1)du \geq 0\,,\end{aligned}$$

and therefore,

$$\int K^2(u)du = \int K_0^2(u)du + 2\int K_0(u)\delta(u)du + \int \delta^2(u)du \geq \int K_0^2(u)du\,,$$

which proves that K_0 is the optimal kernel. □

A.4 Proofs of Lemmas 2 and 3 and Theorem 4

A.4.1 Upper Bound

Let $\hat{m}^{(\nu)}(0)$ be the estimator resulting from a local polynomial fit of order p with the Epanechnikov weight function K_0 and bandwidth h_n. Let $K_{0\nu}^* = K_{\nu,p+1}^{opt}$ be its equivalent kernel, as defined in (7). Then,

$$\begin{aligned}
&\sup_{m\in\mathcal{C}} E\left\{\left(\hat{m}_\nu(0) - m^{(\nu)}(0)\right)^2 | X_1,\cdots,X_n\right\} \\
\leq\ & \Big(\int t^{p+1} K_{\nu,p+1}^{opt}(t)dt \frac{C}{(p+1)!}\Big)^2 h_n^{2(p+1-\nu)} + \int {K_{\nu,p+1}^{opt}}^2(t)dt \frac{\sigma^2(0)}{nf_X(0)} h_n^{-(2\nu+1)} \\
\equiv\ & A_1 h_n^{2(p+1-\nu)} + A_2 h_n^{-(2\nu+1)}.
\end{aligned}$$

The bandwidth that minimizes the above quantity is given by

$$h_n = \left(\frac{(2\nu+1)A_2}{2(p+1-\nu)A_1}\right)^{1/(2p+3)}.$$

With this choice of h_n we obtain via simple algebra that

$$\begin{aligned}
&\sup_{m\in\mathcal{C}} E\left\{\left(\hat{m}_\nu(0) - m^{(\nu)}(0)\right)^2 | X_1,\cdots,X_n\right\} \\
\leq\ & A_1^s A_2^r \left(\left(\frac{2\nu+1}{2(p+1-\nu)}\right)^r + \left(\frac{2\nu+1}{2(p+1-\nu)}\right)^{-s}\right) \\
=\ & A_1^s A_2^r (2p+3)(2\nu+1)^{-s}[2(p+1-\nu)]^{-r} \\
=\ & Dr^{-r}s^{-s}\Big(\int t^{p+1} K_{\nu,p+1}^{opt}(t)dt\Big)^{2s}\Big(\int (K_{\nu,p+1}^{opt}(t))^2 dt\Big)^r, \qquad \text{(A.8)}
\end{aligned}$$

where

$$D = \left(\Big(\frac{C}{(p+1)!}\Big)^{2\nu+1}\Big(\frac{\sigma^2(0)}{nf_X(0)}\Big)^{p+1-\nu}\right)^{\frac{2}{2p+3}}.$$

Letting

$$c_p = \frac{(p+\nu+2)!}{(\frac{p+1+\nu}{2})!\,(\frac{p+1-\nu}{2})!}, \qquad \text{(A.9)}$$

then, according to Gasser, Müller and Mammitzsch (1985),

$$\begin{aligned}
\Big|\int t^{p+1} K_{\nu,p+1}^{opt}(t)dt\Big| &= \frac{c_p((p+1)!)^2}{(2p+3)!} \\
\int (K_{\nu,p+1}^{opt}(t))^2 dt &= \frac{c_p^2(p+1-\nu)^2}{(2\nu+1)(2p+3)(p+\nu+2)2^{2p+2}}. \qquad \text{(A.10)}
\end{aligned}$$

Therefore (A.8) can be further simplified by

$$\begin{aligned} & Dr^{-r}s^{-s}c_p^2\frac{((p+1)!)^{4s}}{[(2p+3)!]^{2s}}\frac{(p+1-\nu)^{2r}}{[(2\nu+1)(2p+3)(p+\nu+2)]^r 2^{2r(p+1)}} \\ = \; & Dc_p^2\frac{r^r[(p+1)!]^{4s}}{s[(2p+3)!]^{2s}(p+\nu+2)^r 4^{r(p+2)}} \\ = \; & r_{\nu,p}, \end{aligned} \tag{A.11}$$

leading to an upper bound for the linear minimax risk, where $r_{\nu,p}$ is defined by (19).

A.4.2 Lower Bound

Let us evaluate the modulus of continuity defined by (15). Take an $f \in \mathcal{C}$ and let

$$m_1(x) = \delta^{p+1}f(x/\delta), \qquad m_0(x) = -m_1(x), \tag{A.12}$$

where δ is a positive constant to be determined later on. Clearly, $m_0, m_1 \in \mathcal{C}$.

Now, selecting δ such that

$$\|m_1 - m_0\|^2 = 4\delta^{2p+3}\|f\|^2 = \varepsilon^2,$$

which is equivalent to taking

$$\delta = \left(\frac{\varepsilon^2}{4\|f\|^2}\right)^{1/(2p+3)},$$

we obtain that

$$\omega_\nu(\varepsilon) \geq |m_1^{(\nu)}(0) - m_0^{(\nu)}(0)| = 2|f^{(\nu)}(0)|\left(\frac{\varepsilon^2}{4\|f\|^2}\right)^{\frac{p+1-\nu}{2p+3}}. \tag{A.13}$$

The optimal kernel of order $(\nu, p+1)$ is given by (see Gasser, Müller and Mammitzsch (1985))

$$K_{\nu,p+1}^{opt}(x) = \sum_{j=0}^{p+1}\lambda_j x^j, \tag{A.14}$$

where

$$\lambda_j = \begin{cases} 0 & \text{if } j+p+1 \text{ odd} \\ (-1)^{(j+\nu)/2}\dfrac{c_p(p+1-\nu)(p+1+j)!}{j!(j+\nu+1)2^{2p+3}\left(\frac{p+1-j}{2}\right)!\left(\frac{p+1+j}{2}\right)!} & \text{if } j+p+1 \text{ even,} \end{cases} \tag{A.15}$$

with c_p as in (A.9).

Defining

$$g(x) = \begin{cases} K^{opt}_{\nu,p+1}(x) & \text{if } |x| \le 1 \\ 0 & \text{otherwise,} \end{cases} \tag{A.16}$$

we need the following lemma.

Lemma 4 *The function $g(\cdot)$ satisfies*

$$|g(x) - \sum_{j=0}^{p} g^{(j)}(0)\frac{x^j}{j!}| \le |\frac{g^{(p+1)}(0)}{(p+1)!}x^{p+1}|. \tag{A.17}$$

Proof. It is obvious that (A.17) holds whenever $|x| \le 1$. When $|x| > 1$, $g(x) = 0$ and (A.17) becomes

$$|\sum_{j=0}^{p} \lambda_j x^j| \le |\lambda_{p+1}x^{p+1}|. \tag{A.18}$$

Since the polynomial above is either even or odd (see (A.15)),we only need to check (A.18) for $x > 1$. Note that the polynomial (A.14) has k roots on $[-1, 1]$ (see Lemmas 2 and 3 of Gasser, Müller and Mammitzsch (1985)). Hence $K^{opt}_{\nu,p+1}(x)$ can not change its sign when $x > 1$. Thus

$$\text{sgn}(\lambda_{p+1})K^{opt}_{\nu,p+1}(x) > 0,$$

namely

$$|\lambda_{p+1}|x^{p+1} > -\text{sgn}(\lambda_{p+1})\sum_{j=0}^{p} \lambda_j x^j \qquad \text{when } x > 1.$$

Using a similar argument we get,

$$-\text{sgn}(\lambda_{p+1})\sum_{j=0}^{p} \lambda_j x^j > 0 \qquad \text{when } x > 1.$$

Combining the above two statements, we obtain (A.18). □

We are now ready to establish a lower bound for the linear minimax risk. Take,

$$f(x) = g(ax) \qquad \text{with } a = \left(\frac{C}{(p+1)!\lambda_{p+1}}\right)^{1/(p+1)}, \tag{A.19}$$

where g is defined by (A.16). Then, by Lemma 1, $f \in \mathcal{C}$. It follows from (A.19) that

$$\|f\|^2 = \frac{\|g\|^2}{a} = \frac{\|K^{opt}_{\nu,p+1}\|^2}{a}, \qquad f^{(\nu)}(0) = a^\nu \nu! \lambda_\nu. \tag{A.20}$$

Substituting (A.20) into (A.13), we obtain

$$\omega_\nu(\varepsilon) \geq 2\nu!|\lambda_\nu| \left(\frac{C}{(p+1)!|\lambda_{p+1}|}\right)^s \|K^{opt}_{\nu,p+1}\|^{-r} \left(\frac{\varepsilon^2}{4}\right)^{r/2}. \tag{A.21}$$

Applying Theorem 6 of Fan (1993), we find that the minimax bound of the best linear procedure is

$$\begin{aligned} R_L(\nu) &\geq r^r s^s \left[\nu!\lambda_\nu \left(\frac{C}{(p+1)!|\lambda_{p+1}|}\right)^s \|K^{opt}_{\nu,p+1}\|^{-r} \left(\frac{\sigma^2(0)}{n f_X(0)}\right)^{r/2}\right]^2 \\ &= D\frac{r^r s^s (\nu!)^2 {\lambda_\nu}^2}{(\lambda_{p+1})^{2s}\|K^{opt}_{\nu,p+1}\|^{2r}}. \end{aligned}$$

By definitions (A.9) and (A.15), this leads to

$$\begin{aligned} & R_L(\nu) \\ &\geq \frac{c_p^2 D r^r s^s r^2}{s^2 2^{4(p+2)}(p+\nu+2)^2} \frac{[(p+1)!]^{4s}(p+\nu+2)^{2s}2^{(2p+4)2s}}{r^{2s}[(2p+3)!]^{2s}} \frac{s^r(p+\nu+2)^r 2^{(2p+4)r}}{r^{2r}} \\ &= c_p^2 D r^{r+2-2s-2r} s^{s-2+r} 2^{(2p+4)(-2+2s+r)}(p+\nu+2)^{-2+2s+r}[(p+1)!]^{4s}[(2p+3)!]^{-2s} \\ &= c_p^2 D r^r s^{-1} 2^{-2(p+2)r}(p+\nu+2)^{-r}[(p+1)!]^{4s}[(2p+3)!]^{-2s} \\ &= r_{\nu,p} \end{aligned} \tag{A.22}$$

leading to a lower bound for the linear minimax risk.

Proof of Lemma 2. Since the lower and upper bound are the same, (A.21) becomes an equality. □

Proof of Lemma 3. Statement (17) follows immediately from the upper and lower bound given in respectively (A.11) and (A.22). The second part of the theorem follows from Theorem 2 above and application of Theorem 6 of Fan (1993). □

Proof of Theorem 4. The maximum risk of $m^{(\nu)}(0)$ is given by (A.11). The result follows from Lemma 3. □

A.5 Proof of Theorem 5

It can easily been shown that the maximum risk of the local linear regression smoother is bounded by the left-hand side of (22). See (10) for a similar expressio Therefore, the minimax risks are bounded by the left-hand side of (22). To establish the lower bound, we apply Theorem 6 of Fan (1993). To this end, let the modulus of continuity be

$$\omega(\varepsilon) = \sup\{|m_1(0) - m_0(0)| : m_0, m_1 \in \mathcal{C}_2, \|m_1 - m_0\| = \varepsilon\},$$

where $\|\cdot\|$ is the L_2-norm. Without loss of generality, we assume that C is a diagonal matrix given by $C = \text{diag}\{\lambda_1, \cdots, \lambda_d\}$. Take

$$m_0(x) = \frac{\delta^2}{2}\left(1 - \delta^{-2}\sum_{j=1}^{d}\lambda_j x_j^2\right)_+^2.$$

Then, $m \in \mathcal{C}_2$. It can easily be computed that

$$\|m_0\|^2 = \frac{2\delta^{4+d}S_d}{|C|^{1/2}d(d+2)(d+4)}.$$

Setting the above expression to $\varepsilon^2/4$ leads to

$$\delta = \left(\frac{|C|^{1/2}d(d+2)(d+4)\varepsilon^2}{8S_d}\right)^{1/(d+4)}.$$

Now, taking the pair $m_1 = -m_0$ and m_0, we have that $\|m_1(0) - m_0(0)\| = \varepsilon$. Therefore,

$$\omega(\varepsilon) \geq |m_1(0) - m_0(0)| = \delta^2 = \left(\frac{|C|^{1/2}d(d+2)(d+4)}{8S_d}\right)^{2/(d+4)}\varepsilon^{4/(d+4)}.$$

Now applying Theorem 6 of Fan (1993) with $p = 4/(d+4)$ and $q = 1-p$, we obtain

$$\begin{aligned} R_L &\geq \frac{p^p q^q}{4}\omega^2\left(2\sqrt{\frac{\sigma^2(0)}{nf_X(0)}}\right)(1+o(1)) \\ &= \frac{d}{4}\left(\frac{2}{S_d}\right)^{\frac{4}{d+4}}(d+2)^{\frac{4}{d+4}}(d+4)^{-\frac{d}{d+4}}\left(\frac{\sigma^4(0)}{n^2 f_X^2(0)}|C|\right)^{\frac{2}{d+4}}(1+o(1)). \quad \text{(A.23)} \end{aligned}$$

The fact that the lower and upper bound are the same lead to the conclusions on linear minimax risk. As for nonlinear minimax risk, the conclusion follows directly from Theorem 6 of Fan (1993) together with (A.23). □

Table 1

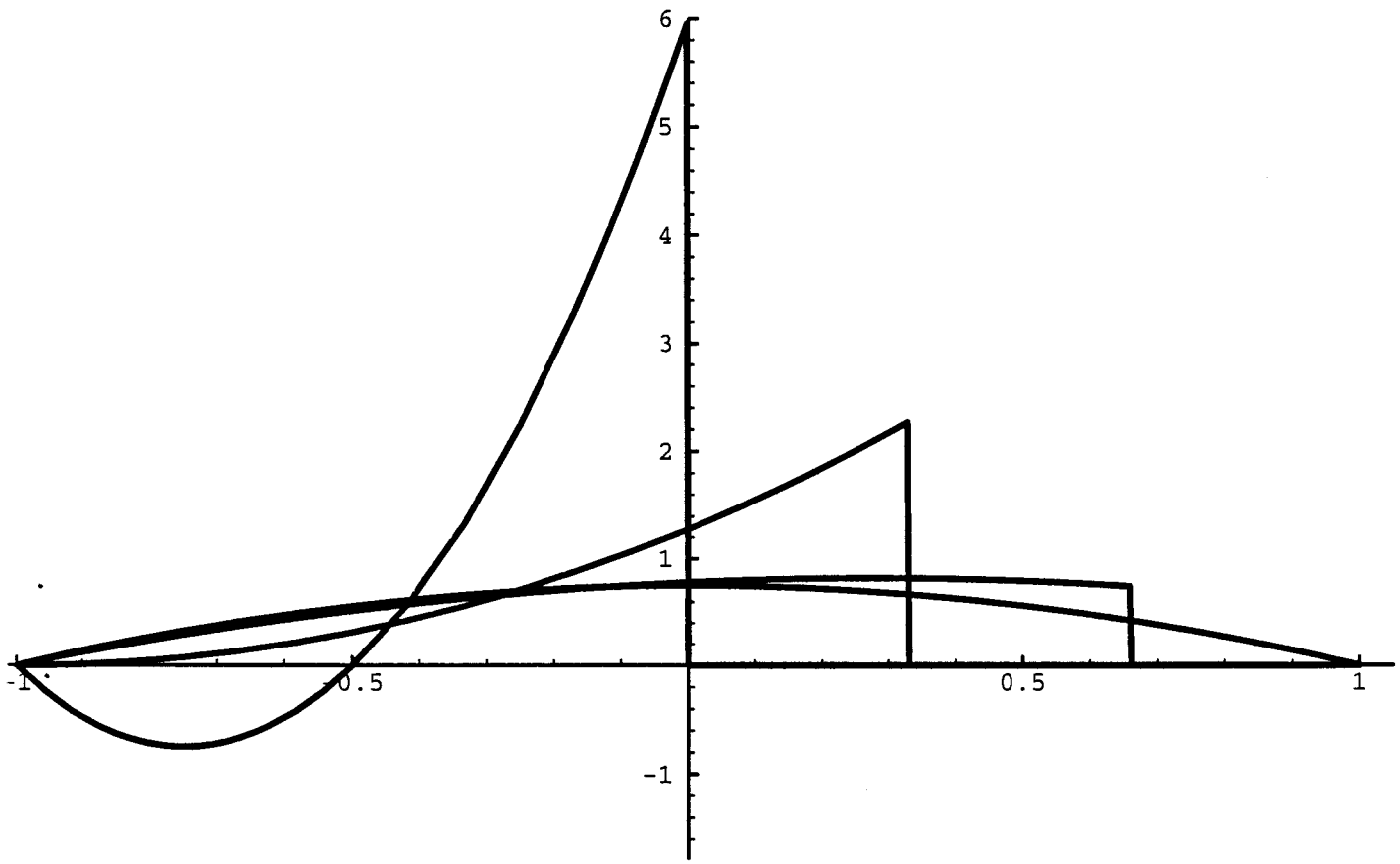
Asymptotically optimal mean integrated square error at the boundary for several boundary kernels of order (ν, k)

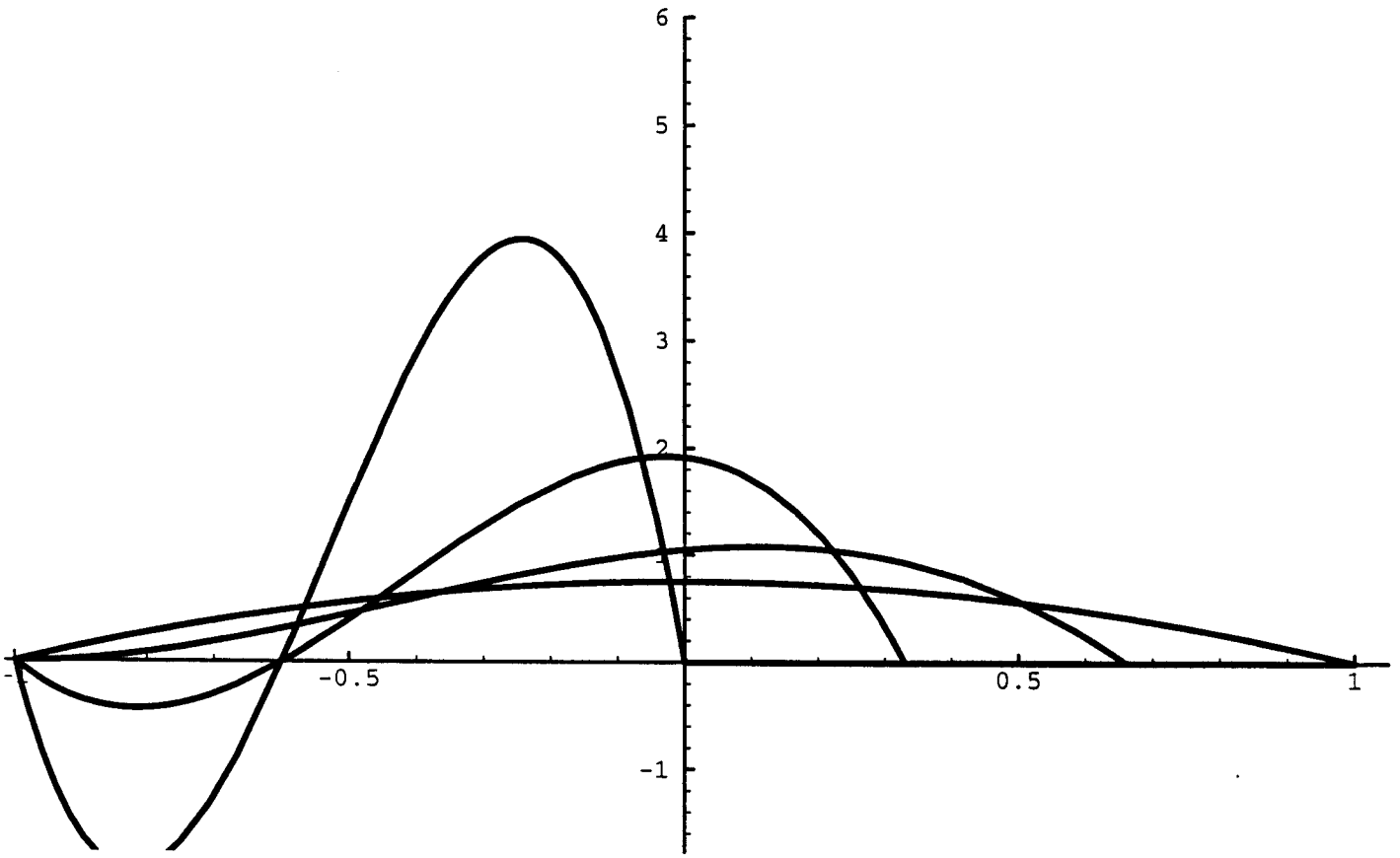
ν	$k = p + 1$	Müller	Minimal order	Equiv. kernel
0	2	0.5616	0.4458	0.4403
0	4	2.017	0.8382	0.8362
1	3	6.242	3.150	3.265
1	5	5349.	1047.	1098.
2	4	2815.	1292.	1357.

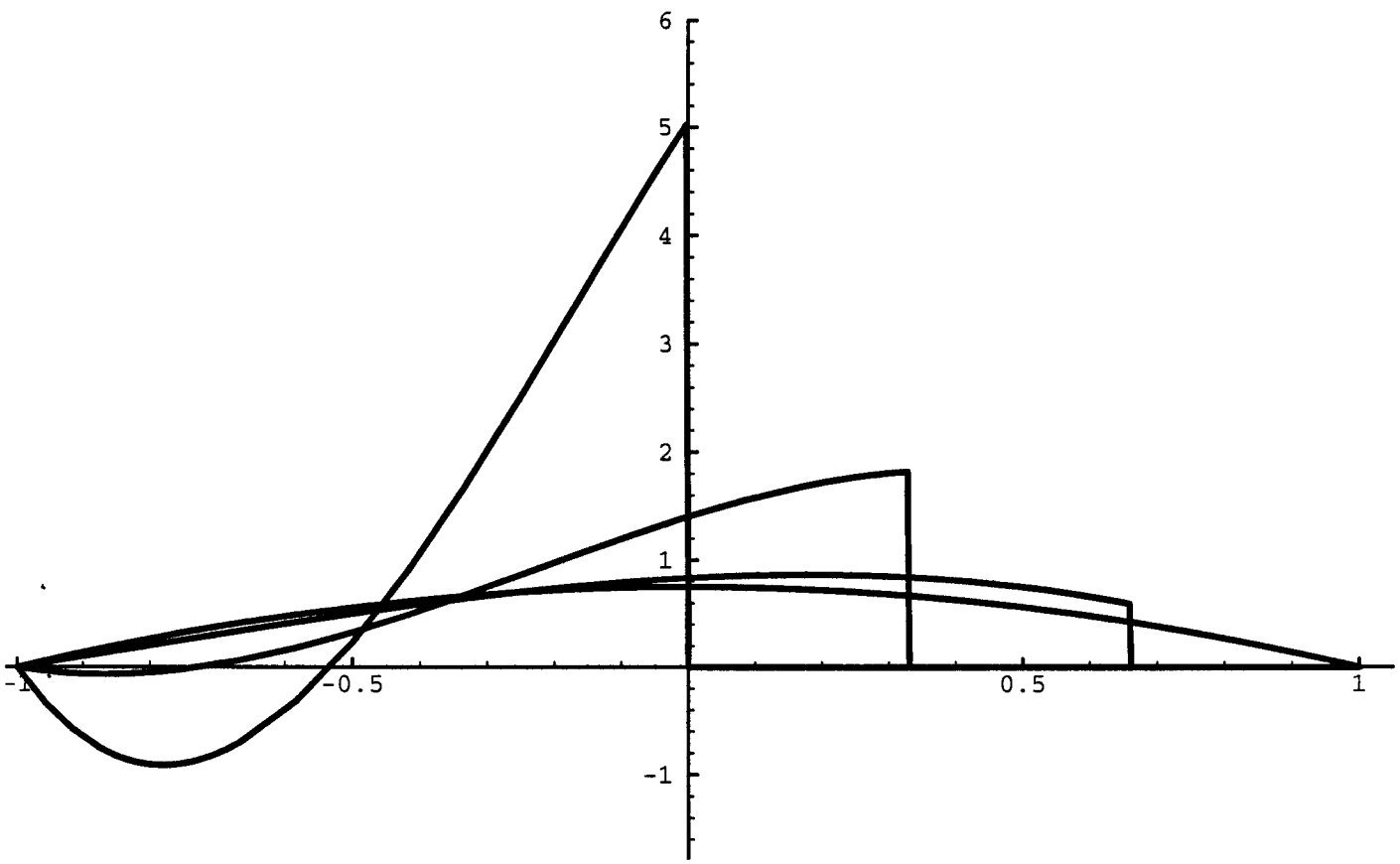
Figure 1: Müller kernel

Figure 2: Minimal order polynomial kernel

Figure 3: Equivalent kernel


6
5
4
3
2
1
-1
0.5
0.5
1


6
5
4
3
2
1
-1
-0.5
0.5
1


6
5
4
3
2
1
-1
-0.5
0.5
1