On Nonparametric Estimation via Local Polynomial 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[3]
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

Abstract. In this paper we focus on local polynomial fitting for estimating a regression function and its derivatives nonparametrically. It has been proven in the literature that this method possesses many nice features, among which automatic adaptation to the boundary and adaptation to various types of designs. A first contribution of this paper is the derivation of an optimal kernel for local polynomial regression at interior points, revealing that their is a universal optimal weighting scheme. Fan (1993) showed that the univariate local linear regression estimator is the best linear smoother in the sense that, with an appropriate choice of the kernel and the bandwidth, this smoother attains the asymptotic linear minimax risk. Moreover, this smoother has high minimax risk. We will show that this property holds also true for the multivariate local linear regression estimator. In the univariate case we investigate minimax efficiency of the more general class of local polynomial regression estimators. It turns out that the asymptotic minimax efficiency for commonly-used orders of fit is 100% among the class of all linear smoothers. Further, we quantify the loss in efficiency when going beyond the class of linear smoothers.

Key words. Curve estimation, local polynomials, minimax efficiency, minimax risk, multivariate curve estimation, nonparametric regression, universal optimal weighting scheme.
Abbreviated title. Local polynomial regression.
AMS 1991 subject classification. Primary 62G07. Secondary 62C20, 62J20.

1 Introduction

In parametric regression the form of the regression function is imposed by the model whereas in nonparametric regression this form is determined by available data. Even when parametric modeling is the ultimate goal, nonparametric methods can prove useful for an exploratory analysis

[1]Supported by the Deutsche Forschungsgemeinschaft.
[2]Supported by NSF grant DMS–9203135.
[3]Supported by the 'Programme d'Action de Recherche Concertée' No. 93-98–164.

and for checking and improving functional models. Details on various nonparametric regression techniques can be found in the monographs of Eubank (1988), Müller (1988), Härdle (1990), Wahba (1990), Green and Silverman (1994) and Wand and Jones (1995), among others.

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 given by

$$m(x) = E(Y|X = x), \tag{1.1}$$

and conditional variance function $\sigma^2(x) = \mathrm{Var}(Y|X = x)$. Non–random $X_1, \ldots, X_n$ are assumed for the following fixed design model:

$$Y_i = m(X_i) + \varepsilon_i \tag{1.2}$$

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 (1.1), a "regular design" is assumed for (1.2), leading formally also to a design density f_X.

Of interest is the estimation of the regression function m and its derivatives. Most of the literature deals with the one–dimensional case $d = 1$. Among linear estimators there are smoothing splines (Wahba (1990)), kernel estimators of the evaluation type (Nadaraya (1964) and Watson (1964)) and of the convolution type (Gasser and Müller (1984)) and local polynomial fitting (for early references see Müller (1988)). Spline smoothers are close to kernel estimators (see Silverman (1984)) with an advantage in terms of minimax properties for the latter (see Jennen–Steinmetz and Gasser (1988)). The relative merits of evaluation and convolution weights for kernel estimation have been discussed by Chu and Marron (1991), Fan (1992), Hastie and Loader (1993) and discussions therein. In summary, the evaluation weights lead to an undesirable form of the bias (Gasser and Engel (1990) and Fan (1992)), while convolution weights pay a price in variance for random designs. Fan (1992) evidenced that the local linear fit overcomes these drawbacks. Moreover, a detailed efficiency study by Fan (1993) revealed that local linear regression estimators (for $d = 1$) achieve full asymptotic minimax efficiency among all linear estimators (for the convolution kernel estimators this holds for fixed design only). Further, the asymptotic minimax efficiency remains at 89.4% among all estimators. Another nice feature of the local linear regression estimator is that the bias at the boundary stays automatically of the same order as in the interior, without use of specific boundary kernels. See Lejeune (1985) and Fan and Gijbels (1992).

In this paper we focus on the more general local polynomial fitting in the univariate ($d = 1$) and the multivariate ($d > 1$) case . Recently, Ruppert and Wand (1994) established asymptotic expressions for the conditional bias and variance of local polynomial regression estimators, and showed that these estimators also adapt automatically to the boundary. The main contribution of the present paper is to investigate the minimax efficiency of local polynomial fits in one and more dimensions. In a first stage, while establishing the framework for this study, we derive a universal optimal weighting scheme under local polynomial fitting.

The paper is organized as follows. Section 2 presents the universal optimal weighting scheme in one dimension, which will be used later on in the efficiency study. Section 3 generalizes the results on the optimal weight function to multivariate local linear fitting. Section 4 contains the main contribution of this paper, i.e. the study on minimax efficiency of local polynomial regression estimators in one and more dimensions.

2 Universal optimal weighting scheme in one dimension

A local polynomial regression at x_0 is computed by minimizing

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} b_j(x_0)\,(X_i - x_0)^j \right\}^2 K\left(\frac{X_i - x_0}{h_n}\right), \tag{2.1}$$

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 (2.1). 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 (2.1) can be written as

$$\begin{aligned} \widehat{b}(x_0) &= (X^T W X)^{-1} X^T W Y \\ &= \begin{pmatrix} S_{n,0}(x_0) & S_{n,1}(x_0) & \cdots & S_{n,p}(x_0) \\ S_{n,1}(x_0) & S_{n,2}(x_0) & \cdots & S_{n,p+1}(x_0) \\ \vdots & \vdots & \ddots & \vdots \\ S_{n,p}(x_0) & S_{n,p+1}(x_0) & \cdots & S_{n,2p}(x_0) \end{pmatrix}^{-1} \begin{pmatrix} T_{n,0}(x_0) \\ T_{n,1}(x_0) \\ \vdots \\ T_{n,p}(x_0) \end{pmatrix} \equiv S_n^{-1} T_n, \end{aligned} \tag{2.2}$$

where

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

Hence

$$\widehat{b}_\nu(x_0) = e_\nu^T \widehat{b}(x_0) = \sum_{i=1}^{n} W_\nu^n\left(\frac{X_i - x_0}{h_n}\right) Y_i \tag{2.3}$$

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

Expression (2.3) reveals that the estimator $\widehat{b}_\nu(x_0)$ is very much like a conventional kernel estimator except that the kernel W_ν^n is defined in terms of the design points X_i and the location point x_0. The weights in (2.3) satisfy the following discrete orthogonality relation:

$$\sum_{i=1}^{n}(X_i - x_0)^q W_\nu^n\left(\frac{X_i - x_0}{h_n}\right) = \delta_{\nu,q} \qquad 0 \le \nu, q \le p, \tag{2.4}$$

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

Calculation of the local polynomial regression estimators reduces to computing the quantities $S_{n,j}(x_0)$ and $T_{n,j}(x_0)$ followed by calculating (2.2). Fast computation algorithms for nonparametric curve estimators, in particular local polynomials, are provided in Fan and Marron (1994) and Seifert *et al.* (1994).

The following lemma provides a representation of the local polynomial regression estimator in terms of an equivalent kernel estimator. Its proof is simple and is omitted.

Lemma 1. *Assume that the design density $f_X(\cdot)$ is continuous and positive at x_0. Then*

$$\widehat{b}_\nu(x_0) = \frac{1}{nh_n^{\nu+1} f_X(x_0)} \sum_{i=1}^{n} K_\nu^*\left(\frac{X_i - x_0}{h_n}\right) Y_i \{1 + o_P(1)\} \tag{2.5}$$

where

$$K_\nu^*(t) = e_\nu^T S^{-1}(1, t, \ldots, t^p)^T K(t) = \sum_{l=0}^{p} S^{\nu l} t^l K(t) \tag{2.6}$$

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

We refer to K_ν^* as the *equivalent kernel.* Since $\int u^q K_\nu^*(u)du = e_\nu^T S^{-1} S e_q$ it follows that the equivalent kernel satisfies the following moment conditions:

$$\int u^q K_\nu^*(u)du = \delta_{\nu,q} \qquad 0 \le \nu, q \le p, \tag{2.7}$$

which are an asymptotic version of the discrete moment conditions presented in (2.4). In fact the equivalent kernel K^*_ν is, up to normalizing constants, a kernel of order $(\nu, p+1)$ as defined by Gasser, Müller and Mammitzsch (1985). Equivalent kernels have been previously used for analyzing polynomial fitting by Lejeune (1985) and Müller (1987) in a slightly different way. See also Ruppert and Wand (1994).

From the above considerations it is easy to derive the asymptotic expressions for the conditional bias and variance of the estimator $\widehat{m}_\nu(x_0)$. As obtained by Ruppert and Wand (1994), these are given by

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

Remark . For polynomial fitting with a symmetric kernel, it is preferable to choose p of the order $p = \nu+1, \nu+3, \ldots$ (see e.g. Ruppert and Wand (1994) and Fan and Gijbels (1995b)). Contrary to expectation, the use of a lower order polynomial $p^* = \nu, \nu+2, \ldots$ with one parameter less 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. See Fan and Gijbels (1995b) for a detailed discussion on the choice of the degree of the polynomial, and for an adaptive procedure for choosing the degree of the polynomial.

Minimization of the asymptotically Mean Squared Error (MSE) leads to the asymptotically optimal local bandwidth

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

Minimizing of the asymptotically Mean Integrated Squared Error (MISE) results in the asymptotically optimal global bandwidth

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

for some weight function $w \geq 0$. It is understood that the denominators do not vanish. Effective estimators of $h_{\mathrm{opt}}(x_0)$ and h_{opt} can be found in Fan and Gijbels (1995a) and Ruppert, Sheather and Wand (1995).

Suppose that Supp(f_X)= $[0,1]$. A point at the left boundary is of the form $x_0 = ch_n$ where $c \geq 0$, and the right boundary points can be defined similarly. The moments are now defined by $s_{j,c} = \int_{-c}^{\infty} u^j K(u)du$ (and $s_{j,c} = \int_{-\infty}^{c} u^j K(u)du$ for the right boundary point $x_0 = 1 - ch_n$), leading to an equivalent boundary kernel (compare with (2.6))

$$K^*_{\nu,c}(t) = e_\nu^T S_c^{-1}(1, t, \ldots, t^p)^T K(t) \quad \text{with } S_c = (s_{j+l,c})_{0 \leq j,l \leq p}. \tag{2.8}$$

This equivalent kernel differs from the one in (2.6) only in the matrix S, and satisfies the boundary moment conditions of Gasser, Müller and Mammitzsch (1985). This reflects clearly that the polynomial method adapts automatically to the boundary, as shown in Fan and Gijbels (1992) for the local linear regression and extended by Ruppert and Wand (1994) to the general case.

Next, the question arises which weight function should be used for different choices of ν and p. The asymptotically MSE and MISE, with optimal choice of the bandwidth, depends on the weight function through

$$T_\nu(K) \equiv \left| \int t^{p+1} K^*_\nu(t)dt \right|^{2\nu+1} \left\{ \int K^{*2}_\nu(t)dt \right\}^{p+1-\nu}. \tag{2.9}$$

"Optimal" kernels K^*_ν , minimizing the right-hand side of (2.9), 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, and reveals a *universal optimal weighting scheme* (i.e. the solution is independent of p and ν) for the interior.

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

The proof of this result is given in Appendix A.1. The optimal kernel K^*_ν, of order $(\nu, p+1)$, is given by (see Gasser, Müller and Mammitzsch (1985))

$$K^{\text{opt}}_{\nu,p}(z) = \sum_{j=0}^{p+1} \lambda_j z^j, \tag{2.10}$$

where

$$\lambda_j = \begin{cases} 0 & \text{if } j+p+1 \text{ odd} \\ \dfrac{(-1)^{(j+\nu)/2} C_p (p+1-\nu)(p+1+j)!}{j!(j+\nu+1)2^{2p+3}(\frac{p+1-j}{2})!(\frac{p+1+j}{2})!} & \text{if } j+p+1 \text{ even,} \end{cases} \tag{2.11}$$

with $C_p = (p+\nu+2)!/\left\{(\frac{p+1+\nu}{2})!(\frac{p+1-\nu}{2})!\right\}$. Moreover, we have the following explicit formulae for its $(p+1)^{th}$ moment and L_2-norm:

$$|\int t^{p+1} K^{\rm opt}_{\nu,p}(t)dt| = \frac{C_p\{(p+1)!\}^2}{(2p+3)!},$$

$$\int \{K^{\rm opt}_{\nu,p}(t)\}^2 dt = \frac{C_p^2(p+1-\nu)^2}{(2\nu+1)(2p+3)(p+\nu+2)2^{2p+2}}. \tag{2.12}$$

Can a similar kind of universal optimal weighting scheme be given at the boundary? For example, if one wants to estimate $m(x_0)$, with $x_0 = 1 - ch_n$ which kernel function should be used? In this case the MSE optimality criterion can be defined as (2.9) but with K^*_ν replaced by $K^*_{\nu,c}$. Denote the resulting criterion by $T^c_{p,\nu}(K)$. When using a local polynomial fit with the Uniform kernel, the weighting scheme still possesses the minimum variance property. The behavior is thus optimal in this sense for the interior as well as for the boundary. This follows immediately from a characterization of minimum variance kernels given by Müller (1991).

The situation is not as simple when taking MSE instead of variance as a criterion for optimality. In the context of kernel estimation there is so far no convincing solution for such optimal boundary kernels available except at the point $c = 0$. This makes it also difficult to judge the quality of the Epanechnikov weight function, when fitting local polynomials at the boundary. For the most left boundary point with $c = 0$, Cheng, Fan and Marron (1993) shows that the triangular kernel $K(z) = (1-z)I_{[0,1]}(z)$ is the optimal kernel, independent of p and ν. (For the most right boundary point, the optimal kernel is $K(z) = (1+z)I_{[-1,0]}(z)$) For other boundary points the solution is not yet obtained. One difficulty is that it is not clear how to define an appropriate target function. For example, the point $x_n = ch_n$ does not correspond to the same point when two different kernel functions are used, each using the optimal bandwidth.

Table 1 shows the value of $T^c_{\nu,p}(K)^{1/(2p+3)}$, which is the constant factor that depends on K in the MSE expression. Recall that the triangular kernel is optimal at the most left boundary point $c = 0$, whereas the Epanechnikov kernel is the optimal one at the least left boundary point $c = 1$ — interior point. Note that for Gaussian kernel, the point $x_n = ch_n$ with $c = 1$ is still a boundary point, whereas for the other kernels, $c = 1$ corresponds to interior points. It is worthwhile to note that the triangular kernel has rather large values of $T^c_{\nu,p}(K)^{1/(2p+3)}$ at interior points ($c = 1$) in comparison with other kernels, since it only uses half of the data. Note also that at the most left boundary point $c = 0$, the Epanechnikov kernel performs very close to the optimal boundary kernel — the triangular kernel. This is a justification to use the Epanechnikov kernel even at boundary regions.

Table 1. *Values of the target function $T^c_{\nu,p}(K)^{1/(2p+3)}$ for several kernels K*

ν	p	Gaussian	Triangular	Uniform	Epanechnikov	Biweight
				$c = 0$		
0	1	2.0731	1.1817	1.2167	1.1856	1.1830
0	3	3.9236	2.0996	2.1387	2.1041	2.1012
1	2	5.2638	3.4588	3.6082	3.4713	3.4630
1	4	19.9751	11.7790	12.2056	11.8194	11.7939
2	3	10.8991	7.7564	8.0845	7.7810	7.7664
				$c = 0.5$		
0	1	0.8010	1.1817	0.6294	0.6155	0.6055
0	3	1.4926	2.0996	0.6284	0.7712	0.7957
1	2	2.3042	3.4588	0.8433	0.9060	0.9161
1	4	3.5218	11.7790	1.4678	0.7734	1.1863
2	3	5.2022	7.7564	0.8313	1.1343	1.265i1
				$c = 1$		
0	1	0.8834	1.1817	0.6084	0.5908	0.5923
0	3	1.3077	2.0996	0.8020	0.7873	0.7893
1	2	0.7293	3.4588	0.9021	0.8647	0.8692
1	4	2.8039	11.7790	1.5257	1.4724	1.4808
2	5	1.7734	7.7564	1.3474	1.2927	1.3010

It is known that the choice of the kernel function K is not very important for the performance of the resulting estimators, both theoretically and empirically. However, since the Epanechnikov kernel is optimal in minimizing MSE and MISE at interior points and is nearly optimal at the most boundary point, we recommend to use this kernel function. From Table 1, one can see that the Biweight kernel function or the Triweight kernel function (not reported in Table 1), which are respectively of form

$$K_2(t) = \text{const}\ \{(1-t^2)_+\}^2 \qquad \text{and} \qquad K_3(t) = \text{const}\ \{(1-t^2)_+\}^3$$

perform very closely to the Epanechnikov kernel. Therefore, these kernels are also recommendable. The structure of this kind of kernels also enables one to implement fast computing algorithms, which is another reason for this recommendation.

3 Local linear fitting in higher dimensions

We consider only multivariate local linear fitting ($p = 1$, $\nu = 0$) which 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. In this section we derive the optimal weighting scheme for multivariate local linear fitting. Minimax properties are discussed in Section 4.

Let K be a function defined in $\mathbb{R}^d$ 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 $X_j = (X_{j1}, \ldots, X_{jd})$, $j = 1, \ldots, n$. Denote by

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

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

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

be the diagonal matrix of weights. Then, the local linear regression is to find $\hat{b}_0(x)$ and $\hat{b}_1(x)$ to minimize

$$\sum_{i=1}^{n} \{Y_i - b_0(x) - b_1^T(x)(X_i - x)\}^2 K_B(X_i - x),$$

and the local linear regression estimator is $\widehat{m}(x) = \widehat{b}_0(x)$. Asymptotic analysis (see Theorem 2.1 of Ruppert and Wand (1994)) yields the following expression for the conditional mean squared error:

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

where $H(x)$ stands for the Hessian matrix of m at x.

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

$$\frac{1}{2}\mu_2^2(K) \operatorname{tr}(HBB^T) H - \frac{\nu_0(K)\sigma^2(x)}{2nf_X(x)|B|}(BB^T)^{-1} = 0$$

(see Rao (1973) p. 72), where $\nu_0(K) = \int K^2(u) du$. If the Hessian matrix $H(x)$ is positive or negative definite this equation has a unique solution

$$BB^T = \left(\frac{\nu_0(K)\sigma^2(x)|H^*|^{1/2}}{\mu_2^2(K) n d f_X(x)}\right)^{\frac{2}{d+4}} (H^*)^{-1} \tag{3.1}$$

with

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

which constitutes a minimum. The optimal bandwidth matrix B itself can be chosen as any matrix satisfying the last equation. As can be seen from (3.2) below, the MSE does not depend on the particular choice of B. Relation (3.1) 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

$$\mathrm{AMSE} = \frac{d+4}{4} d^{\frac{4}{d+4}} \left\{ \nu_0^2(K) \mu_2^d(K) \right\}^{\frac{2}{d+4}} \left\{ \frac{\sigma^4(x)}{n^2 f_X^2(x)} |H^*(x)| \right\}^{\frac{2}{d+4}}. \tag{3.2}$$

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

$$\nu_0^2(K)\mu_2^d(K) = \left\{ \int K^2(u)du \right\}^2 \int u_1^2 K(u)du \cdots \int u_d^2 K(u)du \tag{3.3}$$

is minimized subject to

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

The solution given in the next theorem is derived in Appendix A.2.

Theorem 2. *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. The solution is contrary to a suggestion made by Epanechnikov (1969) to use the product of the one-dimensional kernel of the form $0.75(1 - u_i^2)_+$. With the optimal weight function, we can easily obtain that

$$\mu_2(K_0) = \frac{1}{d+4} \quad \text{and} \quad \nu_0(K_0) = \frac{2d(d+2)}{(d+4)S_d}.$$

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

4 Minimax efficiency

In this section minimax efficiency of univariate and multivariate local polynomial estimators is studied, generalizing results obtained by Fan (1993) for the special case $d = 1$, $p = 1$ and $\nu = 0$. It is shown that the asymptotic minimax efficiency for commonly-used orders of fit is 100% among all linear estimators and 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 and Liu (1991), Donoho (1994) and Donoho, Johnstone, Kerkyacharian and Picard (1995) and the references therein, where attention is focused on density estimation and white noise models.

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

Most commonly-used function estimators are linear, i.e. admit a representation of the form $\sum_{i=1}^{n} w_i(x, X_1, \ldots, X_n)Y_i$. In this section, previous minimax results for univariate local linear fitting (see Fan (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. Consider

$$\mathcal{C}_{p+1} = \{m : |m(z) - \sum_{j=0}^{p} \frac{m^{(j)}(0)}{j!} z^j| \leq C \frac{|z|^{p+1}}{(p+1)!}\}, \tag{4.1}$$

which includes all regression functions whose $(p+1)^{th}$ derivative is bounded by C.

Further basic conditions are needed in order to obtain the minimax results:

<u>Condition A</u>

(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.

The *linear minimax risk* is defined as

$$R_{\nu,L}(n, \mathcal{C}_{p+1}) = \inf_{\hat{S}_\nu \text{ linear}} \sup_{m \in \mathcal{C}_{p+1}} E\left[\left\{\hat{S}_\nu - m^{(\nu)}(0)\right\}^2 |X_1, \cdots, X_n\right], \tag{4.2}$$

and the *minimax risk* is

$$R_\nu(n, \mathcal{C}_{p+1}) = \inf_{\hat{T}_\nu} \sup_{m \in \mathcal{C}_{p+1}} E\left[\left\{\hat{T}_\nu - m^{(\nu)}(0)\right\}^2 |X_1, \cdots, X_n\right]. \tag{4.3}$$

The latter involves all possible estimators including nonparametric estimators with data-driven choice of smoothing parameters.

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

$$\omega_\nu(\varepsilon) = \sup\{|m_1^{(\nu)}(0) - m_0^{(\nu)}(0)| : m_0, m_1 \in \mathcal{C}_{p+1}, \|m_1 - m_0\| = \varepsilon\} \tag{4.4}$$

is a key tool for deriving lower bounds for the minimax risk.

Denote the optimal kernel of order $(\nu, p+1)$ by $K_{\nu,p}^{\rm opt}(x)$ (see (2.10)). Recall that, as shown in Theorem 1, $K_{\nu,p}^{\rm opt}(x) = K_{0\nu}^*(x)$ – the equivalent kernel of the Epanechnikov weight function. Let

$$r = 2(p+1-\nu)/(2p+3), \qquad \text{and} \qquad s = (2\nu+1)/(2p+3), \tag{4.5}$$

and denote

$$\tau_{\nu,p} = r^{-r}s^{-s}\Big\{\frac{C}{(p+1)!}\Big\}^{2s}\Big\{\frac{\sigma^2(0)}{nf_X(0)}\Big\}^r. \tag{4.6}$$

Theorem 3. *The linear minimax risk is bounded from above and below by*

$$B_{\nu,p}\{1+o_P(1)\} \ge R_{\nu,L}(n, \mathcal{C}_{p+1}) \ge b_{\nu,p}\{1+o_P(1)\}, \tag{4.7}$$

where

$$B_{\nu,p} = \tau_{\nu,p}\Big(\int |t^{p+1}K_{\nu,p}^{\rm opt}(t)|dt\Big)^{2s}\Big(\int\{K_{\nu,p}^{\rm opt}(t)\}^2dt\Big)^r$$

and

$$b_{\nu,p} = \tau_{\nu,p}\Big(|\int t^{p+1}K_{\nu,p}^{\rm opt}(t)dt|\Big)^{2s}\Big(\int\{K_{\nu,p}^{\rm opt}(t)\}^2dt\Big)^r.$$

Using expressions in (2.12), it can explicitly be calculated that

$$\begin{aligned} b_{\nu,p} \quad &= \quad \Big(\frac{2p+3}{2\nu+1}\Big)\left\{\frac{(p+\nu+2)!}{\big(\frac{p+1+\nu}{2}\big)!\big(\frac{p+1-\nu}{2}\big)!}\right\}^2\left\{\frac{r}{(p+\nu+2)4^{p+2}}\right\}^r \\ &\times\left\{\frac{(p+1)!C}{(2p+3)!}\right\}^{2s}\left\{\frac{\sigma^2(0)}{nf_X(0)}\right\}^r. \end{aligned} \tag{4.8}$$

How far apart are the lower bound and upper bound in the linear minimax risk? To get an impression of this we present in Table 2 the ratio of the square roots of $b_{\nu,p}$ and $B_{\nu,p}$ (the MSE lower bound and upper bound) which is given by

$$\theta_{\nu,p} = \left(\frac{|\int t^{p+1}K_{\nu,p}^{\rm opt}(t)dt|}{\int |t^{p+1}K_{\nu,p}^{\rm opt}(t)|dt}\right)^s. \tag{4.9}$$

It is easy to see that this ratio is equal to one for the commonly-used cases: $p=1, \nu=0$ and $p=2$, $\nu=1$. We conjecture that the sharp risk is obtained by replacing $|K_{\nu,p}^{\rm opt}|$ in the definition of $B_{\nu,p}$ by $|K_{\nu,p}^{\rm opt*}|$ which minimizes

$$T_{\nu,p}^*(K) = \left\{\int |t^{p+1}K_\nu^*(t)|dt\right\}^{2\nu+1}\left\{\int K_\nu^{*2}(t)dt\right\}^{p+1-\nu}$$

(compare with (2.9)). However, the solution to this problem does not usually admit an explicit formula and has a so complicated shape that it is only of theoretical interest. See Sacks and Ylvisaker (1981).

Table 2. *Ratio $\theta_{\nu,p}$ of the minimax lower and upper bounds*

p $p-\nu$	1	2	3	4	5
1	1	1	.9902	.9124	.7620
3			.9478	.8126	.6611
5					.8983

Let $\widehat{m}_\nu^*(0) = \nu!\widehat{b}_\nu(0)$ be the estimator resulting from a local polynomial fit of order p with the Epanechnikov kernel K_0 and with the bandwidth

$$h_n = \left[\frac{(2\nu+1)(p+1)!^2 \int \left\{K_{\nu,p}^{\mathrm{opt}}(t)\right\}^2 dt\, \sigma^2(0)}{2(p+1-\nu)\left\{\int |t^{p+1}K_{\nu,p}^{\mathrm{opt}}(t)|dt\right\}^2 f_X(0)C^2 n}\right]^{1/(2p+3)}. \tag{4.10}$$

It is worthwhile noting that these choices of the kernel function and the bandwidth reduce to the choices provided in Fan (1993) for the special case that $p = 1$ and $\nu = 0$. With the above choices of the kernel and the bandwidth the local polynomial estimator has a high linear minimax efficiency, as is established in the next theorem.

Theorem 4. *The local polynomial estimator $\widehat{m}_\nu^*(0)$ has high linear minimax efficiency for estimating $m^{(\nu)}(0)$ in the sense that*

$$\frac{R_{\nu,L}(n, \mathcal{C}_{p+1})}{\sup_{m\in\mathcal{C}_{p+1}} E\left[\{\widehat{m}_\nu^*(0) - m^{(\nu)}(0)\}^2 | X_1, \cdots, X_n\right]} \geq \theta_{\nu,p}^2\{1 + o_P(1)\}. \tag{4.11}$$

Thus, for commonly-used local polynomial fitting with $p = 1, \nu = 0$ or $p = 2, \nu = 1$, the estimator $\widehat{m}_\nu^*(0)$ is the best linear estimator. For other values of p and ν (unless the value is too large to be impractical), the local polynomial fit is nearly the best, as evidenced by Table 2. In fact, if our conjecture after Theorem 3 is true, the efficiency of $\widehat{m}_\nu^*(0)$ is far higher than $\theta_{\nu,p}$.

The behavior of the minimax risk $R_\nu(n, \mathcal{C}_{p+1})$ over the class of regression functions $\mathcal{C}_{p+1}$ is established in statement (4.12) below. The second statement in the theorem shows that the local polynomial regression estimator $\widehat{m}_\nu^*(0)$ comes asymptotically fairly close to the minimax risk. This estimator is efficient in the rate, and nearly efficient in the constant factor.

Theorem 5. *The minimax risk (4.3) is asymptotically bounded by*

$$B_{\nu,p} \geq R_\nu(n, \mathcal{C}_{p+1}) \geq (0.894)^2 b_{\nu,p}, \tag{4.12}$$

with $b_{\nu,p}$ and $B_{\nu,p}$ as in (4.7). Moreover, the local polynomial regression estimator $\widehat{m}_\nu^*(0)$ has an asymptotic efficiency of at least 89.4 $\theta_{\nu,p}$ % among all estimators:

$$\frac{R_\nu(n, \mathcal{C}_{p+1})}{\sup_{m \in \mathcal{C}_{p+1}} E\left[\{\widehat{m}_\nu^*(x_0) - m^{(\nu)}(x_0)\}^2 | X_1, \cdots, X_n\right]} \geq (0.894)^2 \theta_{\nu,p}^2 + o_P(1), \tag{4.13}$$

with $\theta_{\nu,p}$ as in (4.9).

The minimax theory provided in this section is an additional justification of the intuitively appealing local polynomial approximation method. Other justifications such as graphical representation and finite sample simulations can be found in Fan (1992), Hastie and Loader (1993) and discussions therein.

4.2 Local linear fitting in higher dimensions

In this section we will show that the multivariate local linear regression estimator, as defined in Section 3, possesses the same minimax properties as the univariate local linear regression estimator: with an appropriate choice of the bandwidth matrix and the kernel function, the multivariate local linear regression estimator achieves asymptotically the linear minimax risk, and comes asymptotically fairly close to the minimax risk. 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 $R_{0,L}(n, \mathcal{C}_2)$ and $R_0(n, \mathcal{C}_2)$ to denote respectively the minimax linear risk and the minimax risk, defined similarly to (4.2) and (4.3). Without loss of generality, we assume that $x = 0$.

Theorem 6. *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 (4.13) in Theorem 5. Moreover, the linear minimax risk is given by*

$$R_{0,L}(n, \mathcal{C}_2) = r_d\{1 + o_P(1)\}, \tag{4.14}$$

where

$$r_d = \frac{d}{4}(\frac{2}{S_d})^{4/(d+4)}(d+2)^{4/(d+4)}(d+4)^{-d/(d+4)} \left\{\frac{\sigma^4(0)}{n^2 f_X^2(0)}|C|\right\}^{2/(d+4)}. \tag{4.15}$$

Further, the minimax risk $R_0(n, \mathcal{C}_2)$ is asymptotically bounded by

$$r_d \geq R_0(n, \mathcal{C}_2) \geq (0.894)^2 r_d. \tag{4.16}$$

In the univariate case (i.e. $d = 1$), the quantity r_d reduces to the factor in Fan (1993).

References

Cheng, M.Y., Fan, J. and Marron, J.S. (1993). Minimax efficiency of local polynomial fit estimators at boundaries. *Institute of Statistics Mimeo Series #2098*, University of North Carolina at Chapel Hill.

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

Donoho, D.L. (1994). Statistical estimation and optimal recovery. *Ann. Statist.*, **22**, 238 – 270.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? *J. Royal Statist. Soc. B*, **57**, 301 – 369.

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.

Epanechnikov, V.A. (1969). Nonparametric estimation of a multidimensional probability density. *Theor. Probab. Appl.*, **13**, 153 – 158.

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. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.*, **20**, 2008 – 2036.

Fan, J. and Gijbels, I. (1995a). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Royal Statist. Soc. B*, **57**, 371 – 394.

Fan, J. and Gijbels, I. (1995b). Adaptive order polynomial fitting: bandwidth robustification and bias reduction. *Journal of Computational and Graphical Statistics*, to appear.

Fan, J. and Marron, J.S. (1994). Fast implementations of nonparametric curve estimators. *Journal of Computational and Graphical Statistics*, **3**, 35 – 56.

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.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach.* Chapman and Hall, London.

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

Hastie, T.J. and Loader, C. (1993). Local regression: automatic kernel carpentry (with discussion). *Statistical Science*, **8**, 120 – 143.

Jennen - Steinmetz, C. and Gasser, T. (1988). A unifying approach to nonparametric regression estimation. *Jour. 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. Appl.*, **9**, 141 – 142.

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

Ruppert, D., Sheather, S.J. and Wand, M.P. (1995). An effective bandwidth selector for local least squares regression. *Jour. Amer. Statist. Assoc.*, to appear.

Ruppert, D. and Wand, M.P. (1994). Multivariate weighted least squares regression. *Ann. Statist.*, **22**, 1346 – 1370.

Sacks, J. and Ylvisaker, D. (1981). Asymptotically optimum kernels for density estimation at a point. *Ann. Statist.*, **9**, 334 – 346.

Seifert, B., Brockmann, M., Engel, J. and Gasser, T. (1994). Fast algorithms for nonparametric curve estimation. *Journal of Computational and Graphical Statistics*, **3**, 192 – 213.

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.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing.* Chapman and Hall, London.

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

Appendix — Proofs of the results

A.1 Proof of Theorem 1

In order to use the same normalization constant as Gasser, Müller and Mammitzsch (1985), we define the equivalent kernel by (compare with (2.6))

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

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 (2.6), 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. $\square$

A.2 Proof of Theorem 2

Let

$$\varphi(K) = \left\{ \int K^2(u)du \right\}^2 \int u_1^2 K(u)du \cdots \int u_d^2 K(u)du \; . \tag{A.1}$$

We had assumed without loss of generality that

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

which leads to

$$\varphi(K) = \left\{ \int K^2(u)du \right\}^2 \left\{ \frac{1}{d} \int (u_1^2 + \cdots + u_d^2) K(u)du \right\}^d \; . \tag{A.3}$$

Property (A.3) allows us to find the optimal K through the following minimization problem:

Minimize $\int K^2(u)du$ subject to

$$\begin{array}{l} \int K(u)du = 1, \quad K \geq 0, \quad \int uK(u)du = 0, \quad \int u_i u_j K(u)du = 0 \quad \text{when } i \neq j \\ \int (u_1^2 + \cdots + u_d^2) K(u)du = \int (u_1^2 + \cdots + u_d^2) K_0(u)du. \end{array} \tag{A.4}$$

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.3 Proofs of Theorems 3 — 5

A.3.1 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 $K_{0\nu}^* = K_{\nu,p}^{\rm opt}$ be its equivalent kernel, as defined in (2.6). Then,

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

The bandwidth that minimizes the above quantity is given by

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

which is exactly the bandwidth defined in (4.10). With this choice of h_n we obtain via simple algebra that

$$\begin{aligned} &\sup_{m \in \mathcal{C}_{p+1}} E\left[\left\{\hat{m}_\nu(0) - m^{(\nu)}(0)\right\}^2 | X_1, \cdots, X_n\right] \\ &\leq A_1^s A_2^r \left[\left\{\frac{2\nu+1}{2(p+1-\nu)}\right\}^r + \left\{\frac{2\nu+1}{2(p+1-\nu)}\right\}^{-s}\right] \\ &= A_1^s A_2^r (2p+3)(2\nu+1)^{-s}[2(p+1-\nu)]^{-r} \\ &= B_{\nu,p}, \end{aligned} \tag{A.5}$$

with $B_{\nu,p}$ as in Theorem 3. This establishes the upper bound for the linear minimax risk.

A.3.2 Lower Bound

Let us evaluate the modulus of continuity defined by (4.4). Take an $f \in \mathcal{C}$ and let

$$m_1(x) = \delta^{p+1} f(x/\delta), \qquad m_0(x) = -m_1(x), \tag{A.6}$$

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

$$\omega_\nu(\varepsilon) \geq |m_1^{(\nu)}(0) - m_0^{(\nu)}(0)| = 2|f^{(\nu)}(0)| \left(\frac{\varepsilon^2}{4\|f\|^2}\right)^{\frac{p+1-\nu}{2p+3}}. \tag{A.7}$$

Recall the optimal kernel of order $(\nu, p+1)$ given in (2.10).

Defining

$$g(x) = \begin{cases} K_{\nu,p}^{\mathrm{opt}}(x) & \text{if } |x| \leq 1 \\ 0 & \text{otherwise,} \end{cases} \tag{A.8}$$

we need the following lemma.

Lemma 2. *The function $g(\cdot)$ satisfies*

$$|g(x) - \sum_{j=0}^{p} g^{(j)}(0)\frac{x^j}{j!}| \leq |\frac{g^{(p+1)}(0)}{(p+1)!}x^{p+1}|. \tag{A.9}$$

Proof. It is obvious that (A.9) holds whenever $|x| \leq 1$. When $|x| > 1$, $g(x) = 0$ and (A.9) becomes

$$|\sum_{j=0}^{p} \lambda_j x^j| \leq |\lambda_{p+1}x^{p+1}|. \tag{A.10}$$

Since the polynomial above is either even or odd (see (2.11)), we only need to check (A.10) for $x > 1$. Note that the polynomial (2.10) has k roots on $[-1, 1]$ (see Lemmas 2 and 3 of Gasser, Müller and Mammitzsch (1985)). Hence $K_{\nu,p}^{\mathrm{opt}}(x)$ can not change its sign when $x > 1$. Thus

$$\mathrm{sgn}(\lambda_{p+1})K_{\nu,p}^{\mathrm{opt}}(x) > 0,$$

namely

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

Using a similar argument we get,

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

Combining the above two statements, we obtain (A.10). $\square$

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

$$f(x) = g(ax) \qquad\qquad \text{with } \ a = \left(\frac{C}{(p+1)!\lambda_{p+1}}\right)^{1/(p+1)}, \tag{A.11}$$

where g is defined by (A.8). Then, by Lemma 1, $f \in \mathcal{C}$. It follows from (A.11) that

$$\|f\|^2 = \frac{\|g\|^2}{a} = \frac{\|K_{\nu,p}^{\text{opt}}\|^2}{a}, \qquad f^{(\nu)}(0) = a^{\nu}\nu!\lambda_{\nu}. \tag{A.12}$$

Substituting (A.12) into (A.7), we obtain

$$\omega_{\nu}(\varepsilon) \geq 2\nu!|\lambda_{\nu}| \left(\frac{C}{(p+1)!|\lambda_{p+1}|}\right)^{s} \|K_{\nu,p}^{\text{opt}}\|^{-r}\left(\frac{\varepsilon^2}{4}\right)^{r/2}. \tag{A.13}$$

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

$$R_{\nu,L}(n,\mathcal{C}_{p+1}) \geq r^r s^s \left[\nu!\lambda_{\nu}\left(\frac{C}{(p+1)!|\lambda_{p+1}|}\right)^{s}\|K_{\nu,p}^{\text{opt}}\|^{-r}\left(\frac{\sigma^2(0)}{nf_X(0)}\right)^{r/2}\right]^2 = b_{\nu,p}, \tag{A.14}$$

using (2.10) and (2.12), with $b_{\nu,p}$ as defined in (4.8). This leads to a lower bound for the linear minimax risk.

Proof of Theorem 3. Statement (4.7) follows immediately from the upper and lower bound given in respectively (A.5) and (A.14). $\square$

Proof of Theorem 4. The maximum risk of $m^{(\nu)}(0)$ is given by (A.5). The result follows from Theorem 3. $\square$

Proof of Theorem 5. Theorem 5 is an immediate consequence of Theorem 3, (A.13) and an application of Theorem 6 of Fan (1993). $\square$

A.4 Proof of Theorem 6

It can easily been shown that the maximum risk of the local linear regression smoother is bounded by the left-hand side of (4.15). See (3.2) for a similar expression. Therefore, the minimax risks are bounded by the left-hand side of (4.15). 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 $C = \operatorname{diag}\{\lambda_1, \cdots, \lambda_d\}$. 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

$$\begin{aligned} R_{0,L}(n, \mathcal{C}_2) &\geq \frac{p^p q^q}{4}\omega^2\left(2\sqrt{\frac{\sigma^2(0)}{nf_X(0)}}\right)(1+o(1)) \\ &= \frac{d}{4}\Big(\frac{2}{S_d}\Big)^{\frac{4}{d+4}}(d+2)^{\frac{4}{d+4}}(d+4)^{-\frac{d}{d+4}}\Big\{\frac{\sigma^4(0)}{n^2 f_X^2(0)}|C|\Big\}^{\frac{2}{d+4}}\{1+o(1)\}. \qquad \text{(A.15)} \end{aligned}$$

The fact that the lower and upper bound are the same lead to the conclusions on linear minimax risk. As for the nonlinear minimax risk $R_{0,L}(n, \mathcal{C}_2)$, the conclusion follows directly from Theorem 6 of Fan (1993) together with (A.15). □