Einbeck:

Multivariate Local Fitting with General Basis Functions

Multivariate Local Fitting with General Basis Functions

Jochen Einbeck[*]
Ludwig Maximilians Universität, Institut für Statistik
Akademiestr. 1, 80799 München, Germany

28th November 2002

Abstract

In this paper we combine the concepts of local smoothing and fitting with basis functions for multivariate predictor variables. We start with arbitrary basis functions and show that the asymptotic variance at interior points is independent of the choice of the basis. Moreover we calculate the asymptotic variance at boundary points. We are not able to compute the asymptotic bias since a Taylor theorem for arbitrary basis functions does not exist. For this reason we focus on basis functions without interactions and derive a Taylor theorem which covers this case. This theorem enables us to calculate the asymptotic bias for interior as well as for boundary points. We demonstrate how advantage can be taken of the idea of local fitting with general basis functions by means of a simulated data set, and also provide a data-driven tool to optimize the basis.

Key Words: Bias reduction, local polynomial fitting, multivariate kernel smoothing, Taylor expansion.

[*] einbeck@stat.uni-muenchen.de

1 Introduction

In the last decades nonparametric smoothing has been one of the most attended and challenging fields in statistics. A widely used concept is that of localizing, where only observations in a neighborhood of the target value are used for the estimation of the regression function.

Nadaraya (1964) and Watson (1964) developed one of the earliest local estimators by simply fitting locally a constant mean value to the data. Stone (1977) was among the first to replace the constant by a line, which reduced the bias of the fit significantly, as Fan (1992) shows. Cleveland (1979) did the next extension and fitted polynomials of arbitrary degree instead of a line. Surprisingly the next step, replacing the polynomial basis $1, x, \ldots, x^p$ by an arbitrary basis $\phi_0(x), \ldots, \phi_p(x)$, as suggested briefly in Ramsay & Silverman (1997), has never been further pursued.

Since local fitting is so far only performed with the polynomial basis, the question of what is special about this particular basis arises. The answer is simple. For this basis Taylor's theorem is available which enables us to interpret the estimated parameters and to calculate the error of the approximation. According to this theorem, whose univariate version was firstly discovered by Brook Taylor (1685-1731) and published 1715 in his book *Methodus incrementorum directa et inversa*, a function m at point x can be approximated by a linear combination of polynomials in a neighborhood of x.

Local fitting with general basis functions will require to find a new Taylor theorem for every basis one wants to use, if some theoretical background is desired. Though this is certainly not possible for every basis, extensions for special cases exist. Einbeck (2001) provides a Taylor theorem covering the case where polynomials are replaced by the powers $\phi(x), \phi^2(x), \ldots, \phi^p(x)$ of an invertible function ϕ. The properties of local modelling with such a power basis are examined, and it is shown that by a suitable basis the results of local polynomial fitting can be significantly improved.

Recently, the general research interest has turned from univariate to multivariate smoothing. Cleveland & Devlin (1988) gave an introduction to multivariate locally weighted regression and showed that the concept is useful in practice.

Further impacts on multivariate local modelling were made by Staniswalis, Messer & Finston (1993), treating kernel estimators for multivariate regression, Wand & Jones (1993), describing bivariate kernel density estimation, and Wand (1992), calculating asymptotic mean square errors for multivariate kernel estimators. In a landmark paper of Ruppert & Wand (1994) asymptotic expressions for bias and variance of the multivariate local linear and quadratic fit are derived.

In Section 2 we will introduce the concept of multivariate local fitting with general basis functions. However, a fully theoretical handling of this estimator is not possible since a Taylor theorem for general basis functions does not exist. In Section 3 we focus on basis functions without interactions. We derive a new Taylor theorem which covers this case and provide asymptotic expressions for bias and variance of the corresponding local estimator. We give an example for fitting with general basis functions by means of a simulated data set in Section 4 and provide a data-driven tool to obtain a suitable basis in Section 5. We finish with the discussion in Section 6.

2 Multivariate locally weighted regression using a general basis

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a set of i.i.d. random variables sampled from a population $(X, Y) \in \mathbb{R}^{d+1}$. Y is a scalar response variable and X a $\mathbb{R}^d$-valued predictor variable with density f having support supp$(f) \in \mathbb{R}^d$. We want to estimate the regression function

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

at a vector $x \in \text{supp}(f)$ nonparametrically, i.e. without assuming m to belong to a parametric family of functions. A model fulfilling (1) is

$$Y_i = m(X_i) + \sigma(X_i)\epsilon, \tag{2}$$

where $\sigma^2(x) = Var(Y|X = x)$ is finite, $E(\epsilon) = \mathbf{0}$, $Var(\epsilon) = \mathbf{I}_d$ and ϵ independent of all $X_i, i = 1, \ldots, n$. Let $\{\phi_j : \mathbb{R}^d \longrightarrow \mathbb{R}, j = 1, \ldots, q\}$ a set of multivariate continuously differentiable basis functions, $\Phi(x) = (\phi_1(x), \ldots, \phi_q(x))^T$ and $\alpha_{1q}(x) := (\alpha_1(x), \ldots, \alpha_q(x))^T$.

The amount of smoothing is determined by a symmetric positive definite bandwidth matrix $H \in \mathbb{R}^{d,d}$. (Often instead of H a nonsingular matrix $B \in \mathbb{R}^{d,d}$ is called bandwidth matrix, where H and B have the relationship $H = BB^T$). Let $K : \mathbb{R}^d \mapsto \mathbb{R}$ be a multivariate kernel function and $K_H(u) = |H|^{-1/2} K(H^{-1/2}u)$. For a detailed description of multivariate kernels and bandwidth matrices see Wand & Jones (1993). The estimator of the function $m(\cdot)$ at point x is $\hat{\alpha}_0(x)$, where $\hat{\alpha}(x) = (\hat{\alpha}_0(x), \hat{\alpha}_{1q}^T(x))^T$ is the minimizer of

$$\sum_{i=1}^{n} \left\{ Y_i - \alpha_0(x) - \alpha_{1q}^T(x)(\Phi(X_i) - \Phi(x)) \right\}^2 K_H(X_i - x). \tag{3}$$

The constant $\Phi(x)$, which only transforms the parameters, is useful because it makes the computation faster and the asymptotic calculations more convenient. This approach covers a wide range of well-known estimators. If, for example, $q = d$, then with $\phi_j(z_1, \ldots, z_d) = z_j$, $j = 1, \ldots, d$ we get the multivariate local linear estimator.

With

$$X_x = \begin{pmatrix} 1 & \phi_1(X_1) - \phi_1(x) & \ldots & \phi_q(X_1) - \phi_q(x) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \phi_1(X_n) - \phi_1(x) & \ldots & \phi_q(X_n) - \phi_q(x) \end{pmatrix},$$

$W_x = \operatorname{diag}(K_H(X_1 - x), \ldots, K_H(X_n - x))$ and $y = (Y_1, \ldots, Y_n)^T$ the least squares problem (3) can be written as

$$min_{\alpha(x)} (y - X_x \alpha(x))^T W_x (y - X_x \alpha(x))$$

and has the solution

$$\hat{\alpha}(x) = (X_x^T W_x X_x)^{-1} X_x^T W_x y, \tag{4}$$

provided that the matrix $X_x^T W_x X_x$ is nonsingular. Thus we obtain

$$\hat{m}(x) = \hat{\alpha}_0(x) = e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x y,$$

where $e_1^T = (1, 0 \ldots, 0) \in \mathbb{R}^{q+1}$. Furthermore,

$$E(\hat{m}(x)|\mathbb{X}) = e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x m, \tag{5}$$

where $m = (m(X_1), \ldots, m(X_n))^T$ and $\mathbb{X} = (X_1, \ldots, X_n)$. Finally the conditional covariance matrix is given by

$$\operatorname{Var}(\hat{m}(x)|\mathbb{X}) = e_1^T (X_x^T W_x X_x)^{-1} (X_x^T \Sigma_x X_x)(X_x^T W_x X_x)^{-1} e_1, \tag{6}$$

where $\Sigma_x = \text{diag}(K_H^2(X_i - x)\sigma^2(X_i))$.

In the following we will provide an asymptotic expression for the variance of the estimator $\hat{m}(x)$. We will treat interior as well as boundary points. Thereby we call a point $x \in \text{supp}(f)$ an interior point if $\{z : H^{-1/2}(x-z) \in \text{supp}(K)\} \subset \text{supp}(f)$; otherwise, x will be called a boundary point. Let

$$\mathcal{D}_{x,H} = \{u : (x + H^{1/2}u) \in \text{supp}(f)\} \cap \text{supp}(K).$$

Then $\mathcal{D}_{x,H} = \text{supp}(K)$ if and only if x is an interior point. Note that we consider x as a fixed point in the case of an interior point, but as a sequence x_n converging sufficiently rapidly to the boundary in the case of a boundary point, ensuring that x is a boundary point for all n (see (A4)). Also let

$$\begin{aligned} M_x &= \int_{\mathcal{D}_{x,H}} \begin{pmatrix} 1 \\ u \end{pmatrix} \begin{pmatrix} 1 & u^T \end{pmatrix} K(u)\,du, \\ N_x &= \int_{\mathcal{D}_{x,H}} \begin{pmatrix} 1 \\ u \end{pmatrix} \begin{pmatrix} 1 & u^T \end{pmatrix} K^2(u)\,du, \\ D_x &= (\nabla\phi_1(x), \dots, \nabla\phi_q(x)), \\ A_{D_x} &= \begin{pmatrix} 1 & \\ & D_x \end{pmatrix}. \end{aligned}$$

The symbol ∇ denotes the gradient function $(\partial_1, \dots, \partial_d)^T = \left(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_d}\right)^T$. Let $\nu_0 = \int K^2(u)\,du$. $o_P(1)$ denotes a sequence of random variables which tends to zero in probability. The asymptotic variance of the estimator $\hat{m}(x)$ is provided by the following theorem:

Theorem 1.

Let x be a fixed element in the interior of supp(f). Then under regularity conditions (A1) to (A3) and (A5)

$$Var(\hat{m}(x)|\mathbb{X}) = \frac{\sigma^2(x)}{nf(x)}|H|^{-1/2}\nu_0(1 + o_P(1)) \tag{7}$$

holds. Let further x be a boundary point, i.e. $x = x_b + H^{1/2}c$, where x_b is a point on the boundary of supp(f) and c is a fixed element of supp(K). Then under conditions (A2) to (A5)

$$\begin{aligned} &Var(\hat{m}(x)|\mathbb{X}) = \qquad (8) \\ &= \frac{\sigma^2(x)}{nf(x)}|H|^{-1/2} e_1^T (A_{D_x}^T M_x A_{D_x})^{-1} A_{D_x}^T N_x A_{D_x} (A_{D_x}^T M_x A_{D_x})^{-1} e_1 (1 + o_P(1)). \end{aligned}$$

Surprisingly, the asymptotical conditional variance of $\hat{m}(x)$ for interior points doesn't depend on the basis function (compare Ruppert & Wand (1994), Theorem 2.1). Thus, with a suitable basis, one could reduce the bias without a rise of the variance. However: For general basis functions we can't compute the asymptotical bias, since a general Taylor theorem is missing. In the next section we will focus on a case where a Taylor theorem is available.

3 Asymptotics for basis functions without interactions

Multivariate locally weighted polynomial regression, described in Ruppert & Wand (1994), is based on the multivariate Taylor theorem, which we will extend in the following. Let $d > 0, p \geq 0$, $U \subset \mathbb{R}^d$ open and U_j the projection of U on the j^{th} coordinate. We impose an invertible basis function $\phi_j \in C^{p+1}(U_j)$ separately on every single coordinate, i.e.

$$\phi_j : U_j \to \mathbb{R}, z_j, \mapsto \phi_j(z_j), j = 1, \ldots, d.$$

For convenience of notation we give the same names to the functions $\phi_j : U \to \mathbb{R}, (z_1, \ldots, z_d) \mapsto \phi_j(z_j)$ picking the j^{th} coordinate. Taking the notation from the previous section, it is $\Phi(z_1, \ldots, z_d) = (\phi_1(z_1), \ldots, \phi_d(z_d))^T$, and the inverse function $\Phi^{-1} : \Phi(U) \to U$ is given by $\Phi^{-1}(z_1, \ldots, z_d) = (\phi_1^{-1}(z_1), \ldots, \phi_d^{-1}(z_d))^T$. The matrix D_x reduces to $P_x = \text{diag}(\phi_j'(x_j))_{1 \leq j \leq d}$. The following theorem holds.

Theorem 2 (Generalized multivariate Taylor expansion).
Assume $U \subset \mathbb{R}^d$ open, $p \geq 0$, $m \in C^{p+1}(U)$, $\Phi : U \to \mathbb{R}^d$ like above, further assume the points x, z and their connection curve $C_S(x, z)$, given by the function $y_\Phi(t) = \Phi^{-1}[\Phi(x) + t(\Phi(z) - \Phi(x))]$, $t \in [0, 1]$, to be in U. Then there exists a point $\zeta \in C_S(x, z)$ with

$$m(z) = m(x) + \sum_{j=1}^{p} \frac{1}{j!} \left[((\Phi(z) - \Phi(x)) \cdot \nabla_\Phi)^j m \right](x) + S_{p+1}(z, x), \qquad (9)$$

where $\nabla_\Phi m(x) = P_x^{-1} \nabla m(x)$, and

$$S_{p+1}(z, x) = \frac{1}{(p+1)!} \left[((\Phi(z) - \Phi(x)) \cdot \nabla_\Phi)^{p+1} m \right](\zeta).$$

For a better understanding and application of this theorem, we set

$$N_m(x) = H_m(x) - P_x^{-1} P_x' \mathrm{diag}(\nabla m(x)),$$

where $H_m(x)$ is the Hessian matrix of m and $P_x' = \mathrm{diag}(\phi_j''(x_j))_{1 \le j \le d}$ the derivative of P_x. Thus $N_m(x)$ equals the Hessian matrix at all entries out of the diagonal, while the diagonal values are modified proportionally to the gradient function of m. Now we can write the generalized Taylor expansion in the form

$$\begin{aligned} m(z) = m(x) \quad &+ \quad (\Phi(z) - \Phi(x))^T P_x^{-1} \nabla m(x) + \\ &+ \quad \frac{1}{2} (\Phi(z) - \Phi(x))^T P_x^{-1} N_m(x) P_x^{-1} (\Phi(z) - \Phi(x)) + S_3(z, x), \end{aligned} \tag{10}$$

which reduces to the usual Taylor theorem by setting $\Phi = \mathrm{id}$.

We denote $x = (x_1, \ldots, x_d)$ and $X_i = (X_{i1}, \ldots, X_{id})$ and work from now on with the design matrix

$$X_x = \begin{pmatrix} 1 & \phi_1(X_{11}) - \phi_1(x_1) & \ldots & \phi_d(X_{1d}) - \phi_d(x_d) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \phi_1(X_{n1}) - \phi_1(x_1) & \ldots & \phi_d(X_{nd}) - \phi_d(x_d) \end{pmatrix},$$

where the ϕ_j are continuously differentiable, but not necessarily invertible. All formulas given from (3) to (6) remain thereby unchanged. Next, we derive asymptotic expressions for bias and variance of $\hat{m}(x)$ at interior as well as boundary points. Let $\mu_2 = \int u^2 K(u)\, du$ and $\nu_0 = \int K^2(u)\, du$. We have the following theorem:

Theorem 3.

Let x be a fixed point in the interior of supp(f). Then under regularity conditions (A1) to (A3) and (A5)

$$Bias(\hat{m}(x)|\mathbb{X}) = \frac{1}{2} \mu_2 tr\,(H N_m(x)) + o_P(tr(H)) \tag{11}$$

and

$$Var(\hat{m}(x)|\mathbb{X}) = \frac{\sigma^2(x)}{n f(x)} |H|^{-1/2} \nu_0 (1 + o_P(1)) \tag{12}$$

hold.

Note that the formula for the conditional bias only differs from the corresponding formular for $\Phi(z) = z$ by using $N_m(x)$ instead of $H_m(x)$ (compare Ruppert & Wand (1994), Theorem 2.1). In the univariate case (11) reduces to

$$\mathrm{Bias}(\hat{m}(x)|\mathbb{X}) = \frac{1}{2} \mu_2 h^2 \left(m''(x) - \frac{\phi''(x)}{\phi'(x)} m'(x) \right) + o_P(h^2). \tag{13}$$

This result gives a hint of how to profit by general basis functions: (13) is minimized for $\phi(x) = m(x)$, thus the bias is reduced if the basis function is as near as possible to the underlying function m. In Sections 4 and 5 we will demonstrate how we can take advantage out of this result.

Now we continue with the treatment of boundary points. The following theorem can be seen as an extension of Theorem 3 which covers the case that the odd-order moments of K (see condition (A1)) do *not* vanish.

Theorem 4.

Let x_b be a point at the boundary of supp(f), $x = x_b + H^{1/2}c$, where c is a fixed element of supp(K). Then under conditions (A2) to (A5)

$$
\begin{aligned}
&Bias(\hat{m}(x)|\mathbb{X}) = \qquad (14)\\
&= \frac{1}{2} e_1^T M_x^{-1} \int_{\mathcal{D}_{x,H}} \begin{pmatrix} 1 \\ u \end{pmatrix} K(u) \{u^T H^{1/2} N_m(x) H^{1/2} u\}\, du + o_P(tr(H))
\end{aligned}
$$

and

$$
Var(\hat{m}(x)|\mathbb{X}) = \frac{\sigma^2(x)}{nf(x)} |H|^{-1/2} \left(e_1^T M_x^{-1} N_x M_x^{-1} e_1 + o_P(1) \right). \qquad (15)
$$

Again the asymptotic bias only differs from the corresponding formula for local linear fitting by the modified Hessian matrix $N_m(x)$. The asymptotic conditional variance at the boundary turns out to be independent of the basis function and is identical to the corresponding formula for a linear basis (see Ruppert & Wand (1994), Theorem 2.2). Note that this result is not self-evident, since we showed in (8) that for arbitrary basis functions the asymptotic variance at the boundary is *not* independent of the basis.

Finally recall that in the beginning of the section we defined the function $\Phi(\cdot)$ to be invertible on U. Here U is a neighborhood of x which becomes arbitrarily small for large n, see condition (A3). Thus it is sufficient if the basis functions are *locally* invertible around the target value x, what is already guaranteed by (A5).

4 Example

In this example we contaminate the underlying function $m : [0,1]^2 \longrightarrow \mathbb{R}$,

$$m(x_1, x_2) = (1 - x_1)\sin(12x_2) + x_1^2 \cos(16x_1) \tag{16}$$

with Gaussian noise ($\sigma = 0.25$). The $n = 961$ design points are uniformly distributed on $[0,1]^2$. The function without and with contamination is shown in Fig. 1. For assessing the quality of the fit we use the relative squared error

$$\mathrm{RSE}(\hat{m}) = \frac{\|\hat{m} - m\|}{\|m\|} = \frac{\sqrt{\sum_{i=1}^{n}(m(X_i) - \hat{m}(X_i))^2}}{\sqrt{\sum_{i=1}^{n} m(X_i)^2}}. \tag{17}$$

In the following table we compare the results of the local fit with various basis functions. For reasons of comparability we restrict on the case of two basis functions. From the first to the last line we will increase the amount of information which we install in the basis. Note that all basis functions fall in the general framework of Section 2, whereas only a) to e) fit the setting of Section 3. We provide the bandwidths h_1 and h_2 which minimize the RSE, and the value of RSE obtained at this minimizing bandwidth.

	$\phi_1(x)$	$\phi_2(x)$	h_1	h_2	RSE
$a)$	x_1	x_2	0.03	0.04	0.162
$b)$	$\cos 16x_1$	x_2	0.14	0.03	0.132
$c)$	$x_1^2 \cos 16x_1$	x_2	0.11	0.03	0.108
$d)$	x_1	$\sin 12x_2$	0.03	1.00	0.080
$e)$	$\cos 16x_1$	$\sin 12x_2$	0.04	1.00	0.059
$f)$	$x_1^2 \cos 16x_1$	$(1 - x_1)\sin 12x_2$	1.00	1.00	0.011
$g)$	$\phi_{0.33,0.15^2}(x_1) \cdot \phi_{0.67,0.15^2}(x_2)$	$\phi_{0.67,0.15^2}(x_1) \cdot \phi_{0.33,0.15^2}(x_2)$	0.04	0.03	0.164

Table 1: Relative squared errors for various basis functions.

$\phi_{\mu,\sigma^2}(\cdot)$ denotes the density of a normal distribution with mean μ and variance σ^2. The bandwidth was restricted to a maximum value of 1. For a), d), f) and g) the results are illustrated in Fig. 2, where the plots of the basis functions and the corresponding fits are shown.

The observations obtained from the table and the figures are the following:

- The more information the basis carries about the underlying function, the better the local fit and the higher the optimal bandwidths, see b) to f).
- If by accident a basis is used which doesn't contain any information about the underlying function, as in g), the results fortunately stay similar like for the linear basis. This result is simply explicable: The linear basis is a *wrong* basis. Mostly it does not contain any information about the true function. Thus replacing a wrong basis with another wrong basis will not make much difference.

Now we have the chance to use given information in an effective way. If one has any notion about the true function one can use this information in the basis. If the basis was more or less correct the fit can be improved tremendously.

5 Finding a data-driven basis function

A logical objection to this methodology will be that usually no information about the true function is available. The question is then whether a data-driven method to obtain a suitable basis exists?

We said that the fit will improve if one uses a basis which is similar to the true function. There is a well-known way to obtain a function which is similar to the true function: *Smoothing*. This gives us the following idea: We perform a simple local linear fit and use the result as a basis function.

Returning to the previous example this means that we use the function in the top right of Fig. 2 as our basis function. Fitting only to this basis we obtain an RSE value of 0.143 (at optimal bandwidths $h_1 = 0.29, h_2 = 0.20$), which is already a good part better than the relative error in a). However, the resulting fit in Fig. 3 (top right) seems to be identical to the basis we used. What is happening? The second step - smoothing with the data-driven basis - does not change the local properties of the basis. If there is a wiggly structure in the basis, this wiggly structure will be retained after the second fit. Nevertheless the fit is improved, because the global properties of the basis are modified. The range of the basis obtained from the fit in a) is $(-1.28, 0.93)$. If we smooth the data with this basis, the range blows up to $(-1.36, 1.06)$, i.e. this smoothing step is in fact a kind of backwards-smoothing of the basis, which is corrects fit where the basis was oversmoothed.

This observation motivates us not to use the optimal bandwidths in the first fit, but somewhat higher bandwidths, in order to avoid a wiggly basis. Of course the result will be oversmoothed, but this will be corrected in the second fit. In our example calculating a local linear fit with $h_1 = 0.06$ and $h_2 = 0.08$ leads to $RSE(\hat{m}) = 0.318$, which is certainly not a very good fit, as shown in Fig. 3 (bottom left). However, if we use this fit as a basis for the second fit the RSE can be optimized down to 0.122, what is an impressive improvement. The resulting smooth curve is shown in Fig. 3 (bottom right).

Summarizing the findings, we suggest the following algorithm (for $d \geq 1$ dimensions).

1. Calculate a d-dimensional local linear fit, using the double size of the optimal bandwidths. The optimal bandwidth matrix $H^{1/2} = \text{diag}(h_i)_{1 \leq i \leq d}$ can be obtained by applying usual multivariate local linear bandwidth selection routines, see e.g. Yang & Tschering (1999).

2. Use the result as a d-dimensional basis for the second fit. As a rule of thumb, use of the same bandwidths as in the first fit leads to satisfactory results.

For the verification of this algorithm we did 200 simulations of the contaminated function (16), and plotted the corresponding RSE for the local linear fit (using the optimal bandwidth) and the fit according to the algorithm in boxplots. Since our intention was to explore the benefit of the use of a pre-fit basis and not the performance of local polynomial bandwidth selection procedures, we used the bandwidth minimizing (17) as optimal bandwidth - keeping in mind that this is certainly not possible for a real data set. The boxplots are shown in Fig. 4. The result is obvious and confirms the algorithm.

Note that for $d > 1$ the multivariate pre-fit basis does not fit in the framework of Section 3. Thus the provided example shows that the idea motivated in Section 3 - to use a basis similar to the underlying function - is useful not only if the basis is free of interactions.

6 Discussion

We finish with some considerations about the properties that basis functions should fulfill in theory and practice. Regarding condition (A5), the theory demands the basis functions to be once or twice differentiable and to have non-vanishing gradients at the target point x. Differentiablility, i.e. smoothness, is also important in practice and is fulfilled by all basis functions given in this paper. In particular, the pre-fit basis in Section 5 will be sufficiently smooth for a large initial bandwidth, as proposed in the algorithm.

The condition of non-vanishing gradients is however purely technical and of little practical relevance. In practice, one is free in the choice of a smooth basis, which might or might not contain interactions, and invertibility is no necessary requirement, neither locally nor globally. The fit at points with vanishing gradients will not have apparent drawbacks compared to a fit where (A5) is fulfilled. However, only in the latter case the asymptotic bias and variance can be calculated. Taking any smooth basis, the corresponding theorems hold for all points with non-vanishing gradients of the basis functions. This will usually be fulfilled for all points except a set of measure zero.

In this paper we showed that the concepts of localization and fitting with basis functions can be combined successfully. However we stress that there exists no optimal basis function which could replace the usual polynomial basis in general. The benefit of the application of alternative basis functions depends on the amount of information which is available about the underlying function. If no information is available, we fortunately still can profit by applying the algorithm introduced in Section 5.

There is still plenty of room for further research. For example, it would be desirable to calculate more accurate estimators for the bandwidths which are used in the algorithm. In particular the factor 2, which we use to derive the initial bandwidth from the optimal bandwidth, probably can be further improved. However, a fully theoretical treatment of the pre-fit algorithm will be extremely difficult, since the basis function in the second fit is now a random variable itself.

Acknowledgements

The author is grateful to Gerhard Tutz (LMU, Dep. of Statistics) for helpful inspirations during this work and Daniel Rost (LMU, Math. Inst.) for contributions concerning the extendibility of Taylor‘s theorem. This work was basically finished while the author passed a term at the University São Paulo. Many thanks especially to Carmen D. S. de André and Júlio M. Singer (USP) for their support in various fields.

This paper is to appear in a special issue of *Computational Statistics* on the Euroworkshop on Statististical Modelling (2001). Many thanks to the referees for a variety of suggestions that enhanced the paper.

Appendix

A Regularity conditions

(A1) The kernel K is bounded with compact support, $\int uu^T K(u)du = \mu_2 \mathbf{I}_d$, where μ_2 is a scalar and $\mathbf{I}_d$ the $d \times d$ identity matrix. In addition, all odd-order moments of K vanish, i.e. $\int u_1^{l_1} \cdots u_d^{l_d} K(u)du = 0$ for all nonnegative integers $l_1, \ldots, l_d$ with an odd sum.

(A2) The point x is $\in$ supp(f). At x, σ^2 is continuous, f is continuously differentiable and all second-order derivatives of m are continuous. Further $f(x) > 0$, $\sigma^2(x) > 0$.

(A3) The sequence of bandwidth matrices $H^{1/2}$ is such that $n^{-1}|H|^{-1/2}$ and each entry of H tends to zero as $n \longrightarrow \infty$.

(A4) For a boundary point x, there exists a value x_b on the boundary of supp(f) with $x = x_b + H^{1/2}c$, where c is a fixed element of supp(K), and a convex set $\mathcal{C}$ with nonnull interior containing x_b such that $\underset{x \in \mathcal{C}}{inf} f(x) > 0$.

(A5) At x, all basis functions are continuously differentiable (for variance expressions in Theorem 1,3,4) resp. twice continuously differentiable (for bias expressions in Theorem 3 and 4). In either case, the point x is nonsingular for all basis functions, i.e. $\nabla \phi_j(x) \neq 0$ for $j = 1, \ldots, q$.

For explanations and interpretations of conditions (A1) to (A4) see Ruppert & Wand (1994).

B Proofs

B.1 Proof of Theorem 1

Let $\mathbf{1}$ be a matrix of appropriate dimension having only entries equal to 1, further let

$$A_H = \begin{pmatrix} 1 & 0 \\ 0 & H^{1/2} \end{pmatrix} \in \mathbb{R}^{d+1,d+1}, \text{ and } A_{\mathbf{1}} = \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{1} \end{pmatrix} \in \mathbb{R}^{d+1,q+1}.$$

Note that for any $u \in \mathbb{R}^d$

$$\Phi(x + H^{1/2}u) - \Phi(x) = D_x H^{1/2} u + o(H^{1/2}\mathbf{1})$$

holds. For interior and boundary points we derive

$$\begin{aligned}
&X_x^T W_x X_x = \\
&= \sum_{i=1}^n K_H(X_i - x) \begin{pmatrix} 1 & (\Phi(X_i) - \Phi(x))^T \\ \Phi(X_i) - \Phi(x) & (\Phi(X_i) - \Phi(x))(\Phi(X_i) - \Phi(x))^T \end{pmatrix} \\
&= n \int_{\{t: H^{-1/2}(t-x) \in \mathcal{D}_{x,H}\}} K_H(t-x) \begin{pmatrix} 1 & (\Phi(t)-\Phi(x))^T \\ \Phi(t)-\Phi(x) & (\Phi(t)-\Phi(x))(\Phi(t)-\Phi(x))^T \end{pmatrix} f(t)\, dt \\
&\quad + n o_P(A_{\mathbf{1}}^T A_H \mathbf{1} A_H A_{\mathbf{1}}) \\
&= n f(x) \int_{\mathcal{D}_{x,H}} K(u) \begin{pmatrix} 1 & u^T H^{1/2} D_x \\ D_x^T H^{1/2} u & D_x^T H^{1/2} u u^T H^{1/2} D_x \end{pmatrix} du \qquad (18) \\
&\quad + n o_P(A_{\mathbf{1}}^T A_H \mathbf{1} A_H A_{\mathbf{1}}) \\
&= n f(x) (A_{D_x}^T A_H M_x A_H A_{D_x} + o_P(A_{\mathbf{1}}^T A_H \mathbf{1} A_H A_{\mathbf{1}})). \qquad (19)
\end{aligned}$$

and analogously

$$X_x^T \Sigma_x X_x = n|H|^{-1/2} f(x) \sigma^2(x) (A_{D_x}^T A_H N_x A_H A_{D_x} + o_P(A_{\mathbf{1}}^T A_H \mathbf{1} A_H A_{\mathbf{1}})). \qquad (20)$$

Substituting (19) and (20) into (6) leads to (8). In the special case of an interior point we have $M_x = \begin{pmatrix} 1 & 0 \\ 0 & \mu_2 \mathbf{I}_d \end{pmatrix}$ and $N_x = \begin{pmatrix} \nu_0 & 0 \\ 0 & \int u u^T K^2(u) du \end{pmatrix}$. Thus (8) reduces to

$$\begin{aligned}
\mathrm{Var}(\hat{m}(x)|\mathbb{X}) &= \frac{\sigma^2(x)}{n f(x)} |H|^{-1/2} e_1^T N_x e_1 (1 + o_P(1)) = \qquad (21) \\
&= \frac{\sigma^2(x)}{n f(x)} |H|^{-1/2} \nu_0 (1 + o_P(1)). \qquad (22)
\end{aligned}$$

B.2 Proof of Theorem 2

We introduce the function $M : [0, 1] \to \mathbb{R}$,

$$M(t) = m(y_\Phi(t)) = m(\Phi^{-1}(\Phi(x) + t(\Phi(z) - \Phi(x))).$$

Then we have $M(0) = m(x)$ and $M(1) = m(z)$. We apply the univariate Taylor theorem on the function $M \in C^{p+1}([0, 1])$ and obtain

$$M(1) = M(0) + M'(0) + \frac{1}{2!}M''(0) + \ldots + \frac{1}{p!}M^{(p)}(0) + r_{p+1}, \tag{23}$$

where

$$r_{p+1} = \frac{1}{(p+1)!}M^{(p+1)}(\tau) \quad (\tau \in [0, 1]).$$

Using the Inverse Function Theorem we obtain

$$y'_\Phi(t) = \left[\frac{1}{\phi'_i\left(y_\Phi(t)_{(i)}\right)}(\phi_i(z_i) - \phi_i(x_i))\right]_{(1\le i\le n)}$$

Repeated application of the chain rule on $M = m \circ y_\Phi$ leads to

$$\begin{aligned}
M'(t) &= \nabla m(y_\Phi(t)) \cdot y'_\Phi(t) = [((\Phi(z) - \Phi(x)) \cdot \nabla_\Phi)m](y_\Phi(t)) \\
M''(t) &= [((\Phi(z) - \Phi(x)) \cdot \nabla_\Phi)^2 m](y_\Phi(t)) \\
&\vdots \\
M^{(n)}(t) &= [((\Phi(z) - \Phi(x)) \cdot \nabla_\Phi)^n m](y_\Phi(t))
\end{aligned}$$

Applying the latter formulas in (23) and substituting $\zeta = y_\Phi(\tau)$ proves the allegation.

B.3 Proof of Theorem 3

The proof is kept shortly since it follows mainly the ideas of the corresponding proof for multivariate local linear fitting, see Ruppert & Wand (1994).

Asymptotic Bias

First note that, applying (10), we have

$$m = X_x \begin{pmatrix} m(x) \\ P_x^{-1}\nabla m(x) \end{pmatrix} + \frac{1}{2}Q_m(x) + S_m(x) \tag{24}$$

with

$$Q_m(x) = \left[(\Phi(X_i) - \Phi(x))^T P_x^{-1} N_m(x) P_x^{-1} (\Phi(X_i) - \Phi(x))\right]_{1 \le i \le n}$$

and $S_m(x) = o(Q_m(x))$. Plugging (24) into (5) shows that

$$\text{Bias}(\hat{m}(x)|\mathbb{X}) = \frac{1}{2} e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x Q_m(x)(1 + o(1)). \tag{25}$$

Let $w_i = K_H(X_i - x)$. Using matrix algebra (see e.g. Fahrmeir & Hamerle (1984)) we derive

$$\begin{aligned}
(X_x^T W_x X_x)^{-1} &= \begin{pmatrix} \sum w_i & \sum w_i (\Phi(X_i) - \Phi(x))^T \\ \sum w_i (\Phi(X_i) - \Phi(x)) & \sum w_i (\Phi(X_i) - \Phi(x))(\Phi(X_i) - \Phi(x))^T \end{pmatrix}^{-1} \\
&= n \begin{pmatrix} f(x) + o_P(1) & o_P(\mathbf{1}^T H^{1/2}) \\ o_P(H^{1/2}\mathbf{1}) & \mu_2 P_x H P_x f(x) + o_P(H) \end{pmatrix}^{-1} \\
&= \frac{1}{n} \begin{pmatrix} \frac{1}{f(x)} + o_P(1) & o_P(\mathbf{1}^T H^{-1/2}) \\ o_P(H^{-1/2}\mathbf{1}) & \frac{1}{\mu_2 f(x)} P_x^{-1} H^{-1} P_x^{-1} + o_P(H^{-1}) \end{pmatrix}
\end{aligned} \tag{26}$$

and

$$\begin{aligned}
&X_x^T W_x Q_m(x) = \\
&= \begin{pmatrix} \sum w_i (\Phi(X_i) - \Phi(x))^T P_x^{-1} N_m(x) P_x^{-1} (\Phi(X_i) - \Phi(x)) \\ \sum w_i \left\{ (\Phi(X_i) - \Phi(x))^T P_x^{-1} N_m(x) P_x^{-1} (\Phi(X_i) - \Phi(x)) \right\} (\Phi(X_i) - \Phi(x)) \end{pmatrix} \\
&= n \begin{pmatrix} \mu_2 f(x) \text{tr}\{H N_m(x)\} + o_P(\text{tr}(H)) \\ O_P(H^{3/2}\mathbf{1}) \end{pmatrix},
\end{aligned} \tag{27}$$

so that substituting (26) and (27) into (25) proves (11).

Asymptotic variance

Similar like above we obtain

$$\begin{aligned}
&X_x^T \Sigma_x X_x = \\
&= \begin{pmatrix} \sum w_i^2 \sigma^2(X_i) & \sum w_i^2 \sigma^2(X_i)(\Phi(X_i) - \Phi(x))^T \\ \sum w_i^2 \sigma^2(X_i)(\Phi(X_i) - \Phi(x)) & \sum w_i^2 \sigma^2(X_i)(\Phi(X_i) - \Phi(x))(\Phi(X_i) - \Phi(x))^T \end{pmatrix} \\
&= n \begin{pmatrix} |H|^{-1/2}(\nu_0 \sigma^2(x) f(x) + o_P(1)) & |H|^{-1/2} \mathbf{1}^T H^{1/2}(1 + o_P(1)) \\ |H|^{-1/2} H^{1/2} \mathbf{1}(1 + o_P(1)) & G(x, H) + o_P(|H|^{-1/2} H) \end{pmatrix},
\end{aligned}$$

where

$$G(x, H) = \left(\int K^2(u) u u^T du \right) |H|^{-1/2} P_x H P_x \sigma^2(x) f(x).$$

Plugging this result and (26) into (6) leads to (12).

B.4 Proof of Theorem 4

Let

$$A_H = \begin{pmatrix} 1 & 0 \\ 0 & H^{1/2} \end{pmatrix}, \quad A_{P_x} = \begin{pmatrix} 1 & 0 \\ 0 & P_x \end{pmatrix}.$$

Asymptotic bias

Note that

$$\begin{aligned} X_x^T W_x X_x &= \\ &= n \int_{\{t: H^{-1/2}(t-x) \in \mathcal{D}_{x,H}\}} K_H(t-x) \begin{pmatrix} 1 & (\Phi(t)-\Phi(x))^T \\ (\Phi(t)-\Phi(x)) & (\Phi(t)-\Phi(x))(\Phi(t)-\Phi(x))^T \end{pmatrix} f(t)\, dt \\ &\quad + n o_P(A_H \mathbf{1} A_H) \\ &= n f(x)(A_{P_x} A_H M_x A_H A_{P_x} + o_P(A_H \mathbf{1} A_H)) \qquad (28) \end{aligned}$$

and, using the first step in (27),

$$\begin{aligned} X_x^T W_x Q_m(x) &= \qquad (29) \\ &= n f(x) \begin{pmatrix} \int_{\mathcal{D}_{x,H}} K(u) u^T H^{1/2} N_m(x) H^{1/2} u\, du + o_P(\mathrm{tr}(H)) \\ P_x H^{1/2} \int_{\mathcal{D}_{x,H}} u K(u) \{u^T H^{1/2} N_m(x) H^{1/2} u\}\, du + o_P(H^{1/2} \mathbf{1} \mathrm{tr}(\mathbf{H})) \end{pmatrix} \end{aligned}$$

hold. Assuming (A4), M_x is nonsingular and we have

$$M_x^{-1} = \begin{pmatrix} \mu_x^{11} & \mu_x^{12} \\ \mu_x^{21} & \mu_x^{22} \end{pmatrix},$$

where $\mu_x^{11} = (\mu_{x,11} - \mu_{x,12}\mu_{x,22}^{-1}\mu_{x,21})^{-1}$, $\mu_x^{12} = -(\mu_{x,12}/\mu_{x,11})\mu_x^{22}$ and $\mu_x^{22} = (\mu_{x,22} - \mu_{x,21}\mu_{x,12}/\mu_{x,11})^{-1}$. Then substituting (28) and (29) into (25) and noticing that

$$e_1^T A_{P_x}^{-1} A_H^{-1} M_x^{-1} A_H^{-1} A_{P_x}^{-1} = \begin{pmatrix} \mu_x^{11} & \mu_x^{12} P_x^{-1} H^{-1/2} \end{pmatrix}$$

yields formula (14).

Asymptotic variance

Similar considerations like in (28) lead to

$$X_x^T W_x^2 X_x = n f(x) |H|^{-1/2} (A_{P_x} A_H N_x A_H A_{P_x} + o_P(A_H \mathbf{1} A_H)). \qquad (30)$$

With (6), (28) and (30) we get

$$\begin{aligned}\mathrm{Var}(\hat{m}(x)|\mathbb{X}) &= e_1^T(X_x^TW_xX_x)^{-1}(X_x^TW_x^2X_x)(X_x^TW_xX_x)^{-1}e_1(\sigma^2(x)+o_P(1)) \\ &= \frac{\sigma^2(x)}{nf(x}|H|^{-1/2}\left(e_1^TM_x^{-1}N_xM_x^{-1}e_1+o_P(1)\right),\end{aligned}$$

what had to be proven.

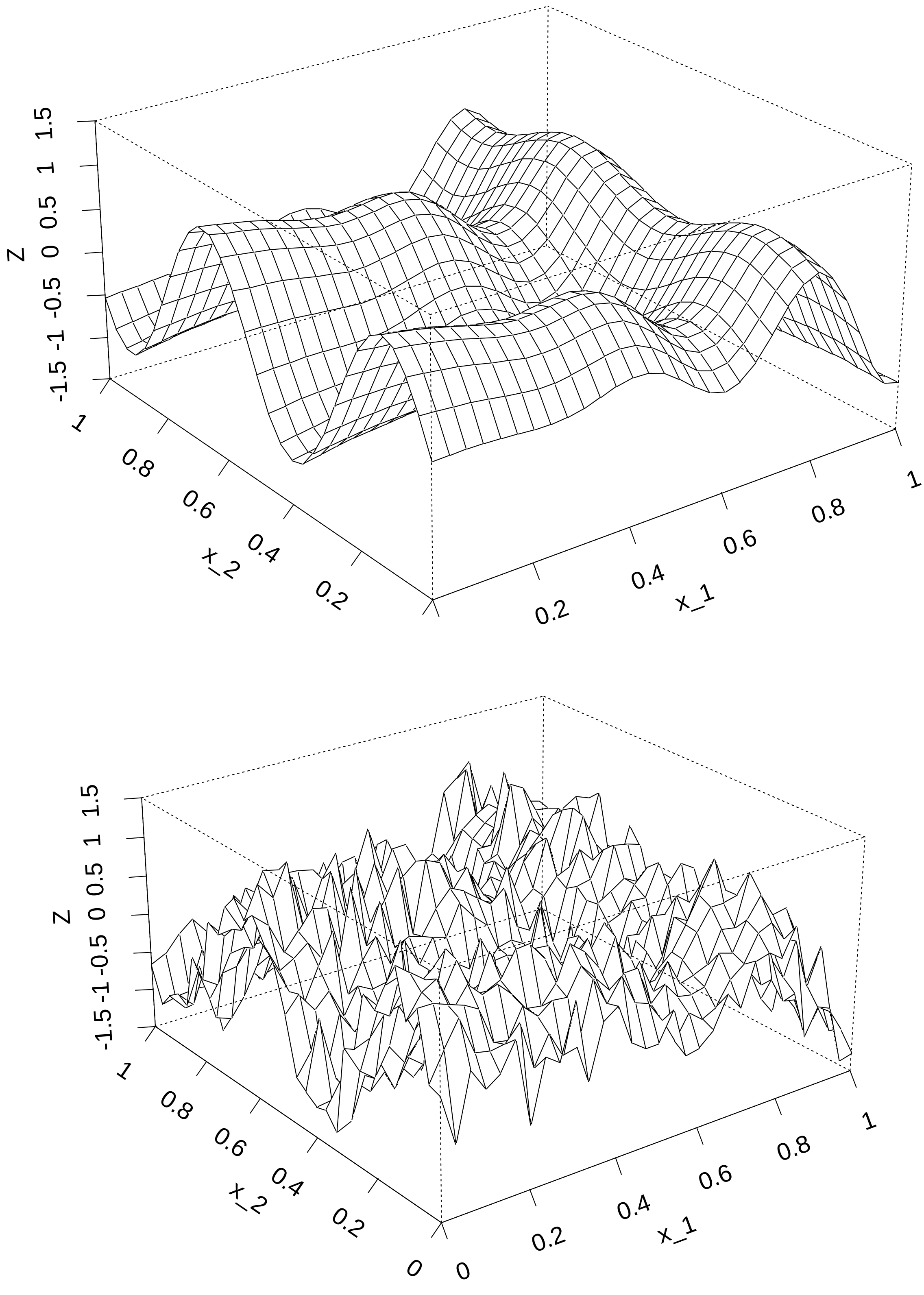

Figure 1: Function (16) without (top) and with contamination (bottom).

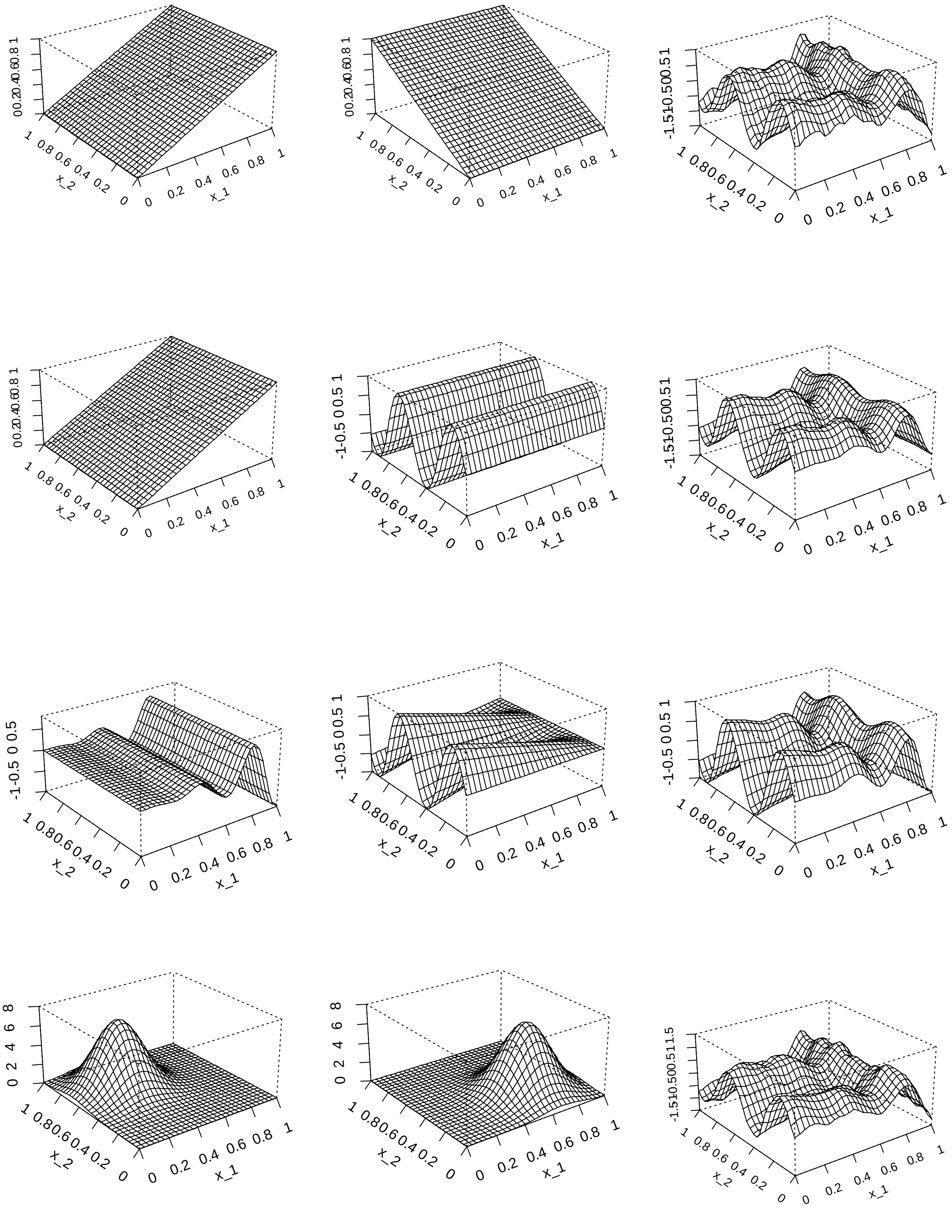

Figure 2: From top to bottom: Basis functions a), d), f) and g) and corresponding fits.

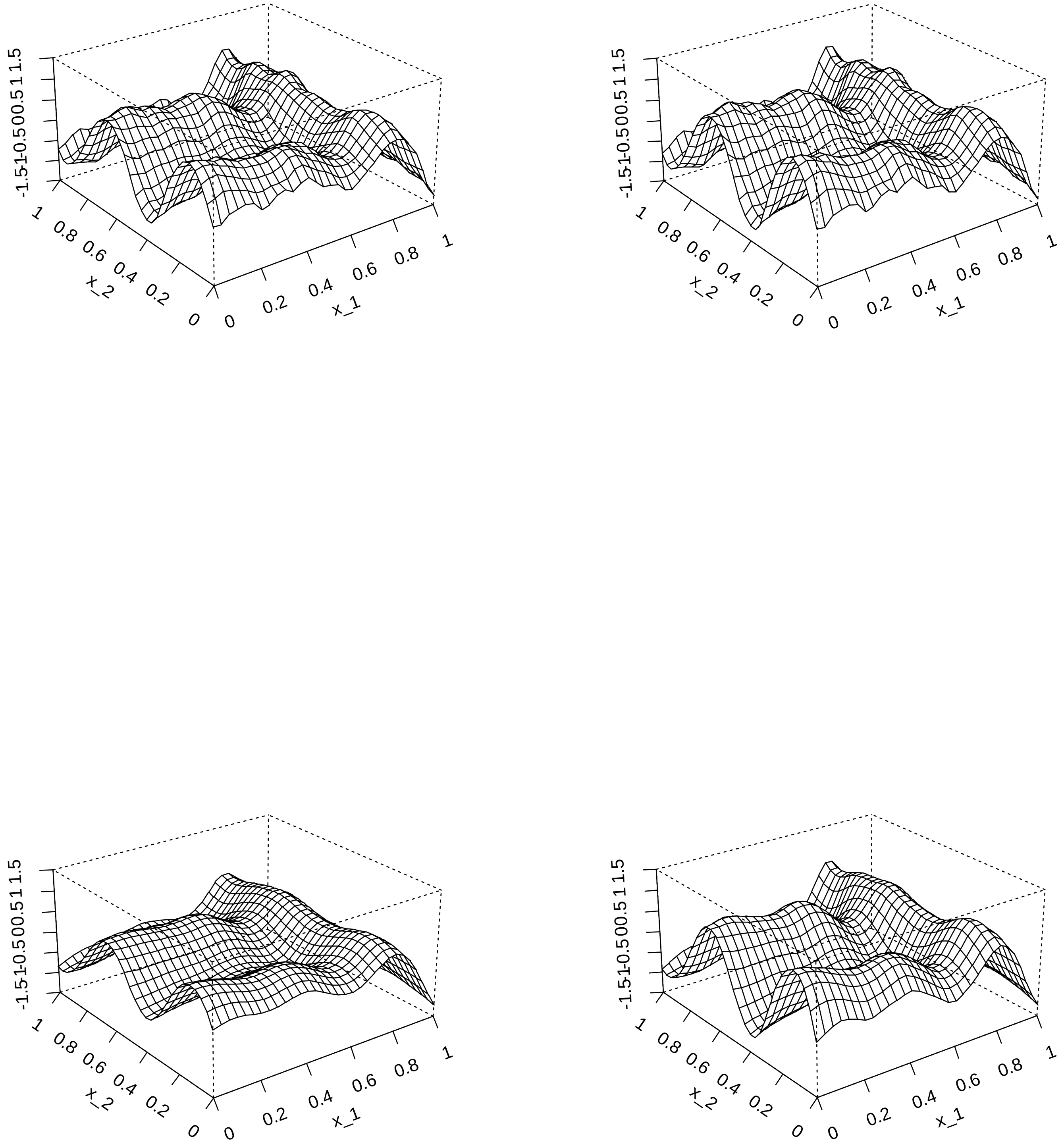

Figure 3: top left: Local linear pre-fit (identical to the top right picture in Fig. 2); bottom left: Local linear pre-fit using the double bandwidth. On the right side is each found the fit obtained using the basis to the left.

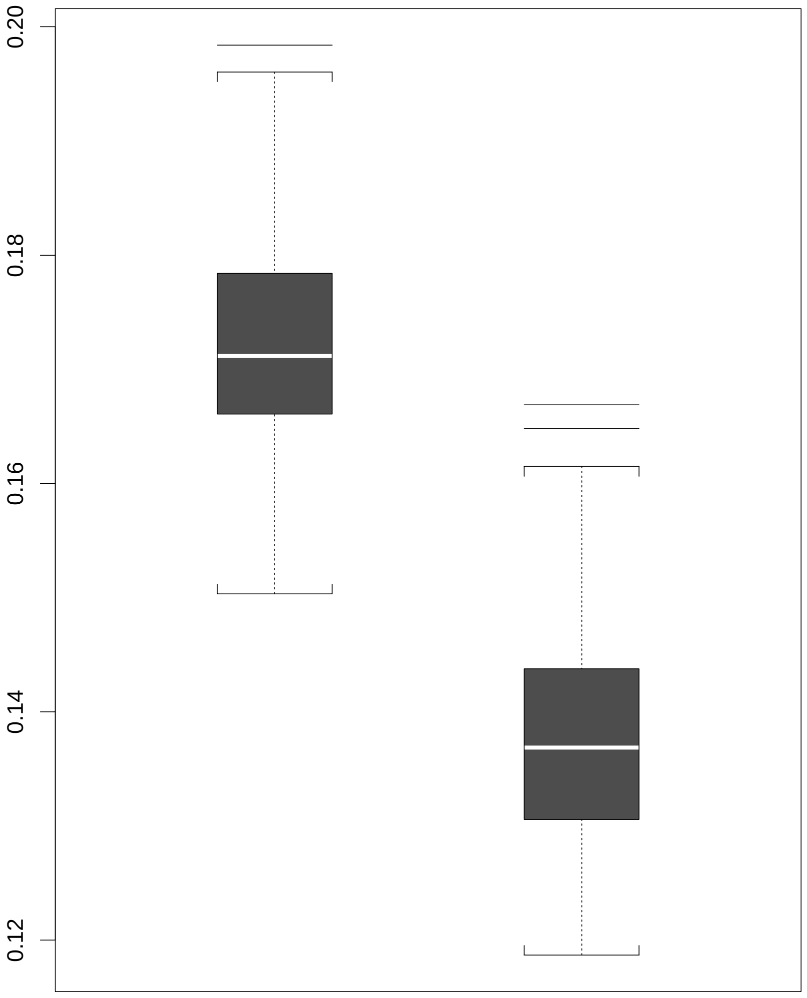

Figure 4: Boxplots of the RSE values of 200 simulations of function (16); left: with local linear basis using the optimal bandwidths; right: with pre-fit basis using the double bandwidths.

References

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

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

Einbeck, J. (2001). Local fitting with general basis functions, SFB 386, Discussion Paper No. 256. *www.stat.uni-muenchen.de/∼einbeck/powerpap06.ps.*

Fahrmeir, L. and Hamerle, A. (1984). *Multivariate statistische Verfahren.* Berlin / New York: de Gruyter.

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

Nadaraya, E. A. (1964). On estimating regression. *Theory Prob. Appl.* **10**, 186–190.

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis.* New York: Springer.

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

Staniswalis, J. G., Messer, K., and Finston, D. R. (1993). Kernel estimators for multivariate regression. *Nonparametric Statistics* **3**, 103–121.

Stone, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5**, 595–645–.

Wand, M. P. (1992). Error analysis for general multivariate kernel estimators. *Nonparametric Statistics* **2**, 1–15.

Wand, M. P. and Jones, M. C. (1993). Comparison of smoothing parametrizations in bivariate kernel density estimation. *J. Amer. Statist. Assoc.* **88**, 520–528.

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

Yang, L. and Tschering, R. (1999). Multivariate bandwidth selection for local linear regression. *J. R. Statist. Soc. B* **61**, 793–815.