Einbeck:

Local Fitting with General Basis Functions

Sonderforschungsbereich 386, Paper 256 (2001)

Local Fitting with General Basis Functions

Jochen Einbeck*
Ludwig Maximilians University
Munich, Germany

3rd August 2001

Abstract

Local polynomial modelling can be seen as a local fit of the data against the basis functions $1, x, \ldots, x^p$. In this paper we extend this method to a wide range of other basis functions. We will focus on the power basis, i.e. a basis which consists of the powers of an arbitrary function, and derive an extended Taylor theorem for this basis. We describe the estimation procedure and calculate asymptotic expressions for bias and variance of this local basis estimator. We apply this method to a simulated data set for various basis functions and propose a data-driven method to find a suitable basis function in each situation.

1 Introduction

In the last years a huge amount of literature about local polynomial modelling has been published. The general framework was given in Fan & Gijbels (1996), followed by various extensions concerning ridging (Seifert & Gasser, 2000), bias reduction (Choi & Hall, 1998), the treatment of measurement errors (Carroll, Maca & Ruppert, 1999 and Lin & Carroll, 2000) and improvements concerning the shape of the smoothing matrix (Zhao, 1999) of the local linear smoother.

Local polynomial modelling is proposed for fitting data points which cannot be modelled satisfactory by global polynomials, like for example the famous

*einbeck@stat.uni-muenchen.de

motorcycle data shown in Fan & Gijbels (1996), p.2. In the following we will shortly review this method.

Consider bivariate data $(X_1, Y_1), \ldots, (X_n, Y_n)$, which form an i.i.d. sample from a population (X, Y). We assume the data to be generated from a model

$$Y = m(X) + \sigma(X)\varepsilon, \tag{1}$$

where $E(\varepsilon) = 0, Var(\varepsilon) = 1$, and X and ε are independent. Of interest is to estimate the regression function $m(x) = E(Y|X = x)$ and its derivatives $m'(x), m''(x), \ldots, m^{(p)}(x)$. A Taylor expansions yields

$$m(z) \approx \sum_{j=0}^{p} \frac{m^{(j)}(x)}{j!}(z-x)^j \equiv \sum_{j=0}^{p} \beta_j(x)(z-x)^j, \tag{2}$$

given that the $(p+1)^{th}$ derivative of $m(\cdot)$ in a neighboorhod of x exists. We define $K_h(\cdot) = \frac{1}{h}K(\frac{\cdot}{h})$, where K is a kernel function which is usually taken to be a non-negative density symmetric about zero. h denotes the bandwidth which determines the size of the neighbourhood of which the covariates which shall influence the fit are chosen. The task of finding the appropriate bandwidth is the crucial point of local polynomial fitting and has been further examined by Ruppert, Sheather & Wand (1995), Hurvich, Simonoff & Tsai (1998) and Doksum, Petersen & Samarov (2000). Minimizing

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j (X_i - x)^j \right\}^2 K_h(X_i - x)$$

leads to the locally weighted least squares regression estimator $\hat{\beta} = (\hat{\beta}_0(x), \ldots, \hat{\beta}_p(x))^T$ and the corresponding estimation functions

$$\hat{m}^{(j)}(x) = j!\hat{\beta}_j(x) \tag{3}$$

for $m^{(j)}(x), j = 0, \ldots, p$. Alternative approaches focussed on estimating the conditional median or quantiles instead of the mean function (see Honda, 2000 and Yu & Jones, 1998), this however will not be of interest in this paper.

According to equation (2), we model data pairs (X, Y) locally around x by

$$Y = \beta_0(x) + \beta_1(x)(X - x) + \ldots + \beta_p(x)(X - x)^p + \sigma(X)\varepsilon.$$

By transforming parameters, this can be written as

$$Y = \alpha_0(x) + \alpha_1(x)X + \ldots + \alpha_p(x)X^p + \sigma(X)\varepsilon.$$

Thus, local polynomial modelling can be interpreted as fitting the data locally against the basis functions $1, X, X^2, \ldots, X^p$.

An obviously arising question is now: Why should just these basis functions be the best possible ones? In this paper we will extend the theory of local polynomial fitting, which is restricted to polynomial basis functions, to a wide range of other basis functions, and give an idea of the advantages and problems coming up by using arbitrary basis functions. A possible choice of basis functions are e.g. Gaussian kernels or the trigonometric functions.

In a general framework one may use the basis functions $\Phi_0(X)$, $\Phi_1(X), \ldots,$ $\Phi_p(X)$, which are arbitrary differentiable functions $\Phi_i : \mathbb{R} \mapsto \mathbb{R}, i = 0, \ldots, p$. This can lead to very good - and tremendously bad - results, as is shown in Section 6. However, theoretical results for the local basis estimator are only available under some restrictions on the basis functions. Regarding (2) and (3), it is seen that estimation is based on Taylor's expansion. In local polynomial fitting, nearly all asymptotic results , e.g. of the bias of the estimator, are based on Taylor's theorem. Asymptotics provide the most important tool to find bandwidth detection rules etc., so that they play an important role for the use of the estimator in practice.

Thus, if we want some theoretical background, we need to develop a new Taylor expansion for every basis we want to use. Of course this will not be possible for all choices of basis functions. In the following section we focus on a special case, namely the power basis, where this is in fact possible and describe the estimation methodology. In Section 3 we provide some asymptotics for estimating the conditional bias and variance of this estimator. In Section 4 we apply this method to a simulated data set and compare the results for various basis functions. We improve these results significantly in Section 5 by using data-adaptive basis functions. In Section 6 we will give an impression of the results obtained when using the most general model.

2 The power basis

Definition 1.
Let $\Phi : \mathbb{R} \longrightarrow \mathbb{R}, x \mapsto \Phi(x)$ *be a differentiable function. Then the functions*

$$1, \Phi(x), \ldots, \Phi^p(x)$$

are called a power basis of degree p.

Taylor's theorem, as found for example in Forster (1999), can be extended as follows:

Theorem 1 (Taylor expansion for a power basis).
Let I *be a non-trivial interval,* $f : I \longrightarrow \mathbb{R}$ *be* $n+1$ *times differentiable in* I *and* $a \in I$*. Then for all* $x \in I$ *a value* $\zeta \in (a, x)$ *resp.* (x, a) *exists so that*

$$f(x) = \sum_{j=0}^{n} \frac{\Psi_{(j)}(a)}{j!}(\Phi(x) - \Phi(a))^j + \frac{\Psi_{(n+1)}(\zeta)}{(n+1)!}(\Phi(x) - \Phi(a))^{(n+1)} \tag{4}$$

with $\Psi_{(j+1)}(x) = \dfrac{\Psi'_{(j)}(x)}{\Phi'(x)}$, $\Psi_0(x) = f(x)$, *holds.*

Assuming the underlying model (1), this theorem suggests to model the data in a neighbourhood of x by

$$Y = \gamma_0(x) + \gamma_1(x)(\Phi(X) - \Phi(x)) + \ldots + \gamma_p(x)(\Phi(X) - \Phi(x))^p + \sigma(X)\varepsilon \tag{5}$$

where

$$\gamma_j(x) = \frac{\Psi_{(j)}(x)}{j!}.$$

One might find the constants $\Phi(x)$ disturbing. However, (5) can easily be transformed to the model

$$Y = \delta_0(x) + \delta_1(x)\Phi(X) + \ldots + \delta_p(x)\Phi^p(X) + \sigma(X)\varepsilon \tag{6}$$

by setting

$$\begin{aligned}\delta_j &= \gamma_j - \gamma_{j+1}\binom{j+1}{1}\Phi(x) + \gamma_{j+2}\binom{j+2}{2}\Phi^2(x) \mp \ldots + \\ &\quad + (-1)^{p-j}\gamma_p\binom{p}{p-j}\Phi^{p-j}(x),\end{aligned}$$

(where for ease of notation $\gamma_j := \gamma_j(x), \delta_j := \delta_j(x)$).

Thus model (5) and (6) yield the same computational results when used for fitting the function m. The advantage of working with model (5) is that its theoretical properties are easier to derive, since the theorem given above can be applied. Moreover, computation is faster and more stable, because very large values are avoided by subtracting $\Phi(x)$ under the powers.

Since the parameters γ_j are constructed more complex than the parameters β_j for local polynomial fitting, the simple relationship $m^{(j)}(x) = j!\beta_j$ can't be retained. However, by using the simple recursive formula

$$\gamma_j(x) = \frac{1}{j\Phi'(x)}\gamma'_{j-1}(x), \quad \gamma_o(x) = m(x),$$

the parameters γ_j can be calculated and thus the following relations between parameters and the underlying function and their derivatives are derived for the power basis:

$$\begin{aligned}
m(x) &= 0!\,\gamma_0 && (7)\\
m'(x) &= 1!\,\Phi'(x)\,\gamma_1 && (8)\\
m''(x) &= 2!\,[\Phi'(x)]^2\,\gamma_2 + \Phi''(x)\,\gamma_1 && (9)\\
m'''(x) &= 3!\,[\Phi'(x)]^3\,\gamma_3 + 3!\,\Phi''(x)\Phi'(x)\,\gamma_2 + \Phi'''(x)\,\gamma_1 && (10)\\
&\vdots
\end{aligned}$$

This indicates that for estimating the ν^{th} derivative of m, the basis function Φ has to be ν times differentiable in an environment of x. It's also possible to calculate these relations for the parameters δ_j, but they are much more difficult, so we omit them here.

In the following we will shortly describe the estimation procedure, which is nearly identical to local polynomial fitting. For ease of comparison, we will use as far as possible the notation introduced in Fan & Gijbels (1996).

In order to estimate $\hat{\gamma} = (\hat{\gamma}_0, \ldots, \hat{\gamma}_p)^T$, a locally weighted least squares regression has to be run, i.e.

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \gamma_j (\Phi(X_i) - \Phi(x))^j \right\}^2 w_i \qquad (11)$$

(with $w_i = K_h(X_i - x)$) has to be minimized in terms of $(\gamma_0, \ldots, \gamma_p)$. The design matrix and the necessary vectors are given by

$$\mathbf{X} = \begin{pmatrix} 1 & \Phi(X_1) - \Phi(x) & \cdots & (\Phi(X_1) - \Phi(x))^p \\ \vdots & \vdots & & \vdots \\ 1 & \Phi(X_n) - \Phi(x) & \cdots & (\Phi(X_n) - \Phi(x))^p \end{pmatrix},$$

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \gamma = \begin{pmatrix} \gamma_0 \\ \vdots \\ \gamma_p \end{pmatrix}, \mathbf{W} = \begin{pmatrix} w_1 & & \\ & \ddots & \\ & & w_n \end{pmatrix}, \mathbf{s} = \begin{pmatrix} \sigma(X_1)\varepsilon \\ \vdots \\ \sigma(X_n)\varepsilon \end{pmatrix}.$$

Then the local fit of (5) corresponds to the fit of

$$\mathbf{y} = \mathbf{X}\gamma + \mathbf{s}$$

and the minimization problem (11) has the form

$$min_\gamma(\mathbf{y} - \mathbf{X}\gamma)^T\mathbf{W}(\mathbf{Y} - \mathbf{X}\gamma), \tag{12}$$

yielding

$$\hat{\gamma} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y},$$

just as in the case of local polynomial fitting. Then

$$\hat{m}(x) = e_1^T\hat{\gamma},$$

where $e_1 = (1, 0, \ldots, 0)^T$, is an estimator for the underlying function $m(\cdot)$ at point x. Using (8) to (10), estimators for the derivatives can be obtained in a similar way. Note that at least p+1 design points are required to lie within the interval $(x - h, x + h)$ to ensure that the matrix $\mathbf{X}^T\mathbf{W}\mathbf{X}$ is invertible.

Furthermore it can be shown that

$$bias(\hat{\gamma}|\mathbb{X}) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{r}, \tag{13}$$

where $\mathbf{r} = (m(X_1), \ldots, m(X_n))^T - \mathbf{X}\gamma$ is the vector of the residuals of the local basis approximation and $\mathbb{X}$ denotes the vector of covariates $(X_1, \ldots, X_n)$. Finally the conditional covariance matrix is given by

$$Var(\hat{\gamma}|\mathbb{X}) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{\Sigma}\ \mathbf{X})(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}, \tag{14}$$

where $\mathbf{\Sigma}$= $diag(w_i^2\sigma^2(X_i))$.

3 Asymptotics

Usually formulas (13) and (14) cannot be used in practice, since they depend on the unknown quantities $\mathbf{r}$ and $\mathbf{\Sigma}$. Consequently an asymptotic derivation is required. In the following, we will go one step further and also approximate all expressions which depend on $X_i, i = 1, \ldots, n$, so that the asymptotic expressions only depend on the location x, but not on the observations.

In the derivations we will use the notations

$$\mu_j = \int_{-\infty}^{\infty} u^j K(u)\, du \quad \text{and} \quad \nu_j = \int_{-\infty}^{\infty} u^j K^2(u)\, du$$

for the j^{th} moments of the kernels K and K^2. Note that $\mu_0 = 1$ and $\mu_{2k+1} = \nu_{2k+1} = 0$ for all $k \in \mathbb{N}_0$. We will omit the integral borders $-\infty$ and ∞ in all following calculations. Further we define the kernel moment matrices

$$\begin{aligned}
&\mathbf{S} = (\mu_{j+l})_{0\le j,l\le p} && \mathbf{c_p} = (\mu_{p+1}, \ldots, \mu_{2p+1})^T\\
&\tilde{\mathbf{S}} = (\mu_{j+l+1})_{0\le j,l\le p} && \tilde{\mathbf{c}}_\mathbf{p} = (\mu_{p+2}, \ldots, \mu_{2p+2})^T\\
&\bar{\mathbf{S}} = ((j+l)\mu_{j+l+1})_{0\le j,l\le p} && \bar{\mathbf{c}}_\mathbf{p} = ((p+1)\mu_{p+2}, \ldots, (2p+2)\mu_{2p+3})^T\\
&\mathbf{S}^* = (\nu_{j+l})_{0\le j,l\le p}\,. &&
\end{aligned}$$

Finally we introduce the denotation $\phi(x) = \Phi'(x)$ and the matrices $\mathbf{H} = diag(h^j)_{0\le j\le p}$ and $\mathbf{P} = diag(\phi^j(x))_{0\le j\le p}$ and recall that $e_{\nu+1} = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ with 1 at $(\nu+1)^{th}$ position. $o_P(1)$ denotes a sequence of random variables which tends to zero in probability, $O_P(1)$ a sequence of random variables which is bounded in probability.

Theorem 2. *Assume that $f(x) > 0$, $\phi(x) \neq 0$ and that $f(\cdot)$, $m^{(p+1)}(\cdot)$, $\Phi^{(p+1)}(\cdot)$ and $\sigma^2(\cdot)$ are continuous in a neighbourhood of x. Further assume that $h \longrightarrow 0$ and $nh \longrightarrow \infty$. Then the asymptotic conditional covariance matrix of $\hat{\gamma}$ is given by*

$$Var(\hat{\gamma}|\mathbb{X}) = \frac{\sigma^2(x)}{nhf(x)} \mathbf{P}^{-1}\mathbf{H}^{-1}\mathbf{S}^{-1}\mathbf{S}^*\mathbf{S}^{-1}\mathbf{H}^{-1}\mathbf{P}^{-1}(1 + o_P(1)). \tag{15}$$

The asymptotic conditional bias is given by

$$bias(\hat{\gamma}|\mathbb{X}) = h^{p+1}\phi^{p+1}(x)\mathbf{P}^{-1}\mathbf{H}^{-1}\left(\gamma_{p+1}\mathbf{S}^{-1}\mathbf{c_p} + b_n)\right), \tag{16}$$

where $b_n = o_P(1)$. *If in addition* $f'(\cdot)$, $m^{(p+2)}(\cdot)$ *and* $\Phi^{(p+2)}(\cdot)$ *are continuous in a neighborhood of* x *and* $nh^3 \longrightarrow \infty$, *the sequence* b_n *can be written as*

$$\begin{aligned} b_n &= h\left[\left(\gamma_{p+1}\frac{f'(x)}{f(x)} + \gamma_{p+2}\phi(x)\right)\mathbf{S}^{-1}\tilde{\mathbf{c}}_\mathbf{p} + \gamma_{p+1}\frac{\phi'(x)}{2\phi(x)}\mathbf{S}^{-1}\bar{\mathbf{c}}_\mathbf{p} - \right. \\ &\quad \left. \gamma_{p+1}\mathbf{S}^{-1}\left(\frac{f'(x)}{f(x)}\tilde{\mathbf{S}} - \frac{\phi'(x)}{2\phi(x)}\bar{\mathbf{S}}\right)\mathbf{S}^{-1}\mathbf{c}_\mathbf{p} + o_P(1) + O_P\left(\frac{1}{\sqrt{nh^3}}\right)\right]. \end{aligned} \tag{17}$$

Based on this theorem and formulas (7) to (10) asymptotic expressions for bias and variance of the mean function and its derivatives can be derived. In particular we obtain for the variance

$$\begin{aligned} Var(\hat{m}(x)|\mathbb{X}) &= Var(e_1^T\hat{\gamma}|\mathbb{X}) \\ &= \frac{\sigma^2(x)}{nhf(x)}e_1^T\mathbf{S}^{-1}\mathbf{S}^*\mathbf{S}^{-1}e_1(1+o_P(1)) \end{aligned}$$

and

$$\begin{aligned} Var(\hat{m}'(x)|\mathbb{X}) &= Var(\phi(x)e_2^T\hat{\gamma}|\mathbb{X}) \\ &= \frac{\sigma^2(x)}{nh^3f(x)}e_2^T\mathbf{S}^{-1}\mathbf{S}^*\mathbf{S}^{-1}e_2(1+o_P(1)). \end{aligned}$$

Thus for $\nu = 0, 1$ the asymptotic variance of $\hat{m}^{(\nu)}(x)$ doesn't depend on the basis function!

Now we take a look at the bias. Using (16) and (7) we arrive at

$$\begin{aligned} bias(\hat{m}(x)|\mathbb{X}) &= bias(e_1^T\hat{\gamma}|\mathbb{X}) \\ &= h^{p+1}\phi^{p+1}(x)e_1^T\left(\frac{\Psi_{(p+1)}(x)}{(p+1)!}\mathbf{S}^{-1}\mathbf{c}_\mathbf{p} + b_n\right) \end{aligned} \tag{18}$$

and

$$\begin{aligned} bias(\hat{m}'(x)|\mathbb{X}) &= bias(\phi(x)e_2^T\hat{\gamma}|\mathbb{X}) \\ &= h^p\phi^{p+1}(x)e_2^T\left(\frac{\Psi_{(p+1)}(x)}{(p+1)!}\mathbf{S}^{-1}\mathbf{c}_\mathbf{p} + b_n\right). \end{aligned}$$

It is important to know that the product $e_{\nu+1}\mathbf{S}^{-1}\mathbf{c}_\mathbf{p}$ is zero for $p-\nu$ even. Thus in the case $p-\nu$ odd, it is sufficient to work with $b_n = o_P(1)$. However, if $p-\nu$ takes an even value, the refined formula (17) for b_n has to be chosen.

Remark: Local Polynomial fitting

Since local fitting based on a power basis is a generalization of local polynomial fitting, setting $\Phi(x) = x$ should correspond to the formulas given by Fan &

Gijbels (1996), Theorem 3.1. Using that for local polynomial fitting $\mathbf{P} = \mathbf{E}$, $\phi = 1$ and $\gamma_j = \beta_j$ holds, we easily find with equation (3)

$$\begin{aligned} Var(\hat{m}^{(\nu)}(x)|\mathbb{X}) &= e_{\nu+1}^T \mathbf{S}^{-1}\mathbf{S}^*\mathbf{S}^{-1} e_{\nu+1} \frac{(\nu!)^2\sigma^2(x)}{f(x)nh^{1+2\nu}} \\ &\quad + o_P\left(\frac{1}{nh^{1+2\nu}}\right) \end{aligned}$$

and

$$\begin{aligned} bias(\hat{m}^{(\nu)}(x)|\mathbb{X}) &= e_{\nu+1}^T \mathbf{S}^{-1}\mathbf{c_p} \frac{\nu!}{(p+1)!} m^{(p+1)}(x) h^{p+1-\nu} \\ &\quad + o_P(h^{p+1-\nu}). \end{aligned}$$

However, in the case $p-\nu$ even, when the more deeply derivation is required, via (16) and (17) we obtain

$$\begin{aligned} bias(\hat{m}^{(\nu)}(x)|\mathbb{X}) &= \nu! h^{p+2-\nu} e_{\nu+1}^T \left[\mathbf{S}^{-1}\tilde{\mathbf{c}}_{\mathbf{p}} \left(\beta_{p+1}\frac{f'(x)}{f(x)} + \beta_{p+2}\right)\right. \\ &\quad \left. - \frac{f'(x)}{f(x)}\beta_{p+1}\mathbf{S}^{-1}\tilde{\mathbf{S}}\mathbf{S}^{-1}\mathbf{c_p}\right] + o_P(h^{p+2-\nu}), \end{aligned}$$

which is not the same as formula (3.9) given in Fan & Gijbels (1996). This difference arises because there on p. 103 it is claimed that the $(\nu+1)^{th}$ element of $e_{\nu+1}\mathbf{S}^{-1}\tilde{\mathbf{S}}\mathbf{S}^{-1}\mathbf{c_p}$ is zero for $p-\nu$ even, which seems to be true for $\nu = 0$, but not in general (consider for e.g. $p = 1$ and $\nu = 1$, then $e_{\nu+1}\mathbf{S}^{-1}\tilde{\mathbf{S}}\mathbf{S}^{-1}\mathbf{c_p} = \mu_2$). The correct formula for general values of $p-\nu$ was provided in Fan, Gijbels, Hu & Huang (1996).

4 Example

Here and in the following sections, we use the L^2 relative error as a measure for the error when estimating the target function:

$$\mathrm{LRE}(\hat{m}) = \frac{\|\hat{m} - m\|}{\|m\|} = \frac{\sqrt{\sum_{i=1}^n (m(X_i) - \hat{m}(X_i))^2}}{\sqrt{\sum_{i=1}^n m(X_i)^2}} \tag{19}$$

As an example, we assume the underlying function

$$m(x) = x + \frac{1}{1.2\sqrt{2\pi}} e^{-(x-0.2)^2/0.02} - \frac{1}{0.9\sqrt{2\pi}} e^{-(x-0.7)^2/0.0018}, \tag{20}$$

which we contaminate with Gaussian noise with $\sigma = 0.05$. Data and function are shown in Fig. 1. The following table shows the "best" values of the relative errors for various basis functions $\Phi(\cdot)$ and degrees p. "Best" means that we choose the bandwidth h_{emp} by

$$h_{emp} = min_h \quad \text{LRE}(\hat{m}).$$

and calculate the relative error for the corresponding value. In order to avoid boundary effects, we omitted the first and last value in the calculation of the relative error.

Φ	$p = 1$	$p = 2$	$p = 3$	$p = 8$
x	0.04878	0.04597	0.04636	0.05116
$\sin x$	0.04862 $*$	0.04649	0.04563 $*$	0.05253
$\arctan x$	0.04861 $**$	0.04656	0.04560 $**$	0.05229
$\text{arcsinh}\, x$	0.04870 $*$	0.04625	0.04599 $*$	0.05038 $*$
x^2	0.04996	0.04572 $*$	0.04920	0.04898 $*$
$\cos x$	0.04991	0.04581 $*$	0.04903	0.04836 $*$
$\cosh x$	0.05000	0.04564 $*$	0.04935	0.04933 $*$
dnorm x	0.04982	0.04601	0.04869	0.04716 $**$
$\exp x$	0.04903	0.04545 $***$	0.04710	0.05014 $*$
$\log(x+1)$	0.04863 $*$	0.04631	0.04600 $*$	0.05170
$\sqrt{x}$	0.04870 $*$	0.04599	0.04772	0.05011 $*$

Table 1: Relative errors for various basis functions and degrees.

In Table 1 dnorm(x) denotes the density of the standard normal distribution. We put a star (*) behind the LRE if the value was better than that for local polynomial fitting, two stars for the winner of the column, and three stars for the over-all-winner. The first four basis functions (included the linear function) are odd functions, followed by four even functions and three more functions without any symmetric properties. The table is easy to read - it shows that odd basis functions perform better for odd values of p and even functions perform better for even values of p. The non-symmetric functions show inhomogeneous behaviour: The exponentional function seems to behave like an even function and provides the over-all-winner, while the logarithmic basis performs like an odd basis. Except for large degrees p, the Gaussian kernels turn out to be not a very good basis. The best choice for odd degrees was the arctan function in all simulations we did. Since the traditional linear basis $\Phi(x) = x$ is an odd function, this is still a good choice when an odd degree is used. If one however wants to fit with an even number of basis functions, it seems to be better to choose an even basis like the cosh function. When doing calculations for the above table, we made another interesting observation: Calculations based on general basis functions were often faster and more stable than for the case of the polynomial basis. First simulations showed that this effect also occurs in the multivariate case, where local polynomial fitting is known to work not very well because of numerical instability. This topic is supposed to be examined in further work.

5 Bias reduction with data-adaptive basis functions

The results in the previous chapter give only very week inducement to change from the polynomial to another basis function. Thus further work is required to find the optimal basis function, what will be done in this section.

Regarding formula (18) in the special case $p = 1$ leads to

$$bias(\hat{m}(x)|\mathbb{X}) = \frac{h^2\mu_2}{2}\left(m''(x) - \frac{\phi'(x)}{\phi(x)}m'(x)\right) + o_P(h^2), \tag{21}$$

what reduces to the well-known formula

$$bias(\hat{m}(x)|\mathbb{X}) = \frac{h^2\mu_2}{2}m''(x) + o_P(h^2),$$

in the special case of local linear fitting. Thus the subtraction of $\frac{\phi'(x)}{\phi(x)}m'(x)$ in (21) provides the chance to bias reduction. In the optimal case, the content of the bracket in (21) is zero, hence the the differential equation

$$m''(x)\phi(x) - m'(x)\phi'(x) = 0$$

has to be solved, what leads to the solutions

$$\phi(x) = c_1 m'(x) \quad (c_1 \in \mathbb{R})$$

and hence

$$\Phi(x) = c_1 m(x) + c_2 \quad (c_1, c_2 \in \mathbb{R}).$$

In particular, for $c_1 = 1, c_2 = 0$ we get $\Phi_{opt}(x) = m(x)$, thus the underlying function $m(\cdot)$ is the optimal basis function. Although the function $m(\cdot)$ is always unknown there are several ways to use this result. For example, one can calculate a pilot estimate $\check{m}(\cdot)$ by doing a local linear fit (or any other smooth estimation, e.g. with splines) and use the estimated function as an improved basis. Or perhaps, in another situation, you may have a notion of the underlying function or know it only partly. For example, let us assume somebody gave us a (wrong) information about the underlying function (20), namely

$$\tilde{m}(x) = x - \frac{1}{1.2\sqrt{2\pi}} e^{-(x-0.2)^2/0.02} - \frac{1}{0.9\sqrt{2\pi}} e^{-(x-0.7)^2/0.0018},$$

i.e. the first hump is showing down instead of up. Then you can try to use this function as basis function.

Applying these approaches to the data set of the previous section leads to the following table. We tried two different bandwiths for the pilot estimate $\check{m}(x)$, once using the best bandwith $h_{emp} = 0.020$, leading to the estimated function $\check{m}_{20}(x)$, and once for a higher bandwith $h = 0.038$, resulting in the fit $\check{m}_{38}(x)$. For comparison we added the results for the linear basis $\Phi(x) = x$ and the optimal basis $\Phi(x) = m(x)$.

Φ	$p=1$	$p=2$	$p=3$	$p=8$
x	0.04878	0.04597	0.04636	0.05116
$\check{m}_{20}(x)$	0.04407 $*$	0.04310 $*$	0.04499 $*$	0.05237
$\check{m}_{38}(x)$	0.03811 $*$	0.03656 $**$	0.03730 $**$	0.05202
$\tilde{m}(x)$	0.03380 $***$	0.03998 $*$	0.04628 $*$	0.05593
$m(x)$	0.00856	0.02587	0.02600	0.03719

Table 2: Relative errors for improved basis functions.

For illustration, we did 50 simulations of the contaminated function (20) , calculated the best relative errors for each the linear, the arctan and the cosh basis as well as for $\tilde{m}(x)$ and $\check{m}(x)$ for $p = 1$ and plotted the corresponding relative errors in boxplots, shown in Fig. 2. The boxplots show that the effect of the basis functions used in Table 1 is negligible, but the performance of the estimation can be improved significantly by using a data-adaptive basis $\check{m}(x)$ (taking each the best bandwidth in the pilot estimate) or the "guessed" basis $\tilde{m}(x)$. The application of the (in praxis unavailable) optimal basis $m(x)$ for $p = 1$ leads to a nearly perfect fit. For higher degrees, results get worse again since the basis was optimized for $p = 1$.

Remark: Bandwidth selection

When the data-adaptive basis function selection procedure is used, two times a bandwidth has to be chosen: one for the pilot estimate, and one for the fit with the data-adaptive basis. For the first fit usual local polynomial bandwidth selection procedures can be applied, see e.g. Fan & Gijbels (1996), p. 110 ff, and Doksum, Petersen & Samarov (2000). If the optimal bandwidth is used for the first fit, then bandwidth selection is not very crucial in the second fit, since this fit is more a correction of the first fit than a localization. In this case the optimal (2^{nd}) bandwidths are very high (in our example $h_{emp} = 0.190$ for p=1) and the minimas are very flat. Hence every large bandwidth will do a good job. This is illustrated in Fig. 3.

However, it is not necessary to find the optimal bandwidth in the first fit. Our simulations showed that the results can even be improved when the optimal bandwith is *not* met, i.e. somewhat higher bandwidths are used. This is seen in Table 2.

If other basis functions, e.g. $\Phi(x) = \arctan x$ or $\Phi(x) = \tilde{m}(x)$, are used, more sophisticated bandwidth selction procedures have to be applied. Using Theorem 2, plug-in formulas for bandwidth selection can be derived straightforward by extending the corresponding methods for local polynomial fitting. This topic however would burst the framework of this paper.

6 Outlook

In Section 1 we already mentioned that the most general model for local basis fitting is

$$Y = \alpha_0(x)\Phi_0(X) + \alpha_1(x)\Phi_1(X) + \ldots + \alpha_p(x)\Phi_p(X) + \sigma(X)\varepsilon. \qquad (22)$$

However, theoretical properties of this fit can (yet) only be analyzed insufficiently. Nevertheless, as will be shown in this section, it is worth to investigate this approach, because the results are somewhat impressive. We analyze the same data set as in Section 4 and 5. We choose $p = 2$ and use $\Phi_0(x) = 1$ and two Gaussian kernels as basis functions $\Phi_{1,2}(x)$. There are two parameters to be adjusted: The distance d between the centers of the kernels and their width, i.e. the standard deviation σ. Taking simply the arbitrary values $\sigma = 0.4$ and $d = 0.35$ yields the relative error 0.04598 for the minimizing bandwidth $h_{emp} = 0.033$, which is a worse fit than those for the polynomials. However, the variation of d and σ leads to significant variations of the L^2 relative error, like shown in Fig. 5 for the bandwidth $h = 0.033$ suggested above. There is an absolute minimum at $\sigma = 0.04$ and $d = 0.39$, yielding a relative error of 0.03132. Using the best bandwidth for this basis function, $h_{emp} = 0.032$ leads to the relative error 0.03128. Again it is obvious that general basis functions can yield much better results than the local polynomial fit, compare also the two pictures in top of Fig. 4. The local basis fit is smoother and closer to the underlying function, especially in the area of the cusps. In the right bottom the optimal basis functions, namely $\Phi_1(x) = \frac{1}{0.04\sqrt{2\pi}} \exp(-(x - 0.305)^2/(2 \cdot 0.04^2))$ and $\Phi_2(x) = \frac{1}{0.04\sqrt{2\pi}} \exp(-(x - 0.695)^2/(2 \cdot 0.04^2))$ are shown. An interesting observation is that the centers of the Gaussian kernels are situated near the humps of $m(\cdot)$, what is simply explicable because we showed in the previous chapter that the better the basis functions model the underlying function, the better are the results.

Unfortunately, the minimum of the goodness-of-fit in Fig. 5 is impossible to find without the knowledge of the underlying function. Nevertheless, it is an interesting observation that the structure of the mountain landscape is more or less similar for all underlying functions $m(\cdot)$. The diagonal ridge results from the singularity of the matrix $\mathbf{X}^T\mathbf{W}\mathbf{X}$ for $d = 2\sigma, d \gtrsim 0.4$, and mostly some craters are situated along the line $d = 2\sigma, d \lesssim 0.4$. However the positions of the other craters seem to be an arbitrary property of the data set and can can be everywhere in or between the mountain ridges. But possibly further research can yield deeper results concerning this problem.

Acknowledgements

The author is grateful to Prof. Dr. Gerhard Tutz, University of Munich, for the fruitful inspirations and discussions, and to Torsten Scholz, University of Munich, for valuable remarks.

Appendix

A Proof of Theorem 1

Define

$$\begin{aligned} g(y) &= f(x) - f(y) - \Psi_{(1)}(y)(\Phi(x) - \Phi(y)) - \frac{\Psi_{(2)}(y)}{2!}(\Phi(x) - \Phi(y))^2 \\ &- \ldots - \frac{\Psi_{(n)}(y)}{n!}(\Phi(x) - \Phi(y))^n - \frac{M}{(n+1)!}(\Phi(x) - \Phi(y))^{n+1}, \end{aligned} \tag{23}$$

where $M \in \mathbb{R}$ is chosen so that $g(a) = 0$ is fulfilled.

Using $g(a) = g(x) = 0$ Rolles theorem (see Forster (1999), S. 153) yields that there exists a $\zeta \in (a, x)$ resp. (x, a) with $g'(\zeta) = 0$. Since

$$g'(y) = -\frac{\Psi'_{(n)}(y)}{n!}(\Phi(x) - \Phi(y))^n - \frac{M}{n!}(\Phi(x) - \Phi(y))^n(-\Phi'(y))$$

it follows that

$$0 = -\Psi'_{(n)}(\zeta) + M\Phi'(\zeta)$$

and thus $M = \Psi_{(n+1)}(\zeta)$. The theorem is obtained by setting $y = a$ in (23).

B Proof of Theorem 2

I. Asymptotic conditional variance

We introduce the notations $S_{n,j} = \sum_{i=1}^{n} w_i(\Phi(X_i) - \Phi(x))^j$ and $S^*_{n,j} = \sum_{i=1}^{n} w_i \sigma^2(X_i)(\Phi(X_i) - \Phi(x))^j$. Then $\mathbf{S_n} := (S_{n,j+l})_{0\le j,l\le p} = \mathbf{X}^T\mathbf{W}\mathbf{X}$ and $\mathbf{S^*_n} := (S^*_{n,j+l})_{0\le j,l\le p} = \mathbf{X}^T\mathbf{\Sigma}\ \mathbf{X}$ hold, and the conditional variance (14) can be written as

$$Var(\hat{\gamma}|\mathbb{X}) = \mathbf{S_n}^{-1}\mathbf{S^*_n}\mathbf{S_n}^{-1} \tag{24}$$

and thus approximation of the matrizes $\mathbf{S_n}$ and $\mathbf{S^*_n}$ is required. Using that

$$\int K(u)u^j g(x+hu)\,du = \mu_j g(x) + o(1) \tag{25}$$

for any function $g : \mathbb{R} \mapsto \mathbb{R}$ which is continuous in x, we obtain

$$\begin{aligned} ES_{n,j} &= n\int K(u)(\Phi(x+hu) - \Phi(x))^j f(x+hu)\,du = \\ &= nh^j \int K(u)u^j\phi^j(\zeta_u)f(x+hu)\,du \\ &= nh^j(f(x)\phi^j(x)\mu_j + o(1)) \end{aligned} \tag{26}$$

where $\zeta_u \in (x, x+hu)$ exists according to Taylor's theorem. Similar we derive

$$\begin{aligned} VarS_{n,j} &= nE(w_1^2(\Phi(X_1) - \Phi(x))^{2j} - nE^2(w_1(\Phi(X_1) - \Phi(x))^j) \\ &= nh^{2j-1}(f(x)\phi^{2j}(x)\nu_{2j} + o(1)) - nh^{2j}(f(x)\phi^j(x)\mu_j + o(1))^2 \\ &= nh^{2j-1}(f(x)\phi^{2j}(x)\nu_{2j} + o(1)) \\ &= n^2h^{2j}O\left(\frac{1}{nh}\right) &\quad (27) \\ &= o(n^2h^{2j}). &\quad (28) \end{aligned}$$

Since for every sequence $(Y_n)_{n\in\mathbb{N}}$ of random variables

$$Y_n = EY_n + O_P\left(\sqrt{VarY_n}\right) \tag{29}$$

holds (what can be proofed with Chebychev's inequality), we can proceed with calculating

$$\begin{aligned} S_{n,j} &= ES_{n_j} + O_P\left(\sqrt{VarS_{n,j}}\right) \\ &= nh^j f(x)\phi^j(x)\mu_j(1 + o_P(1)) \end{aligned} \tag{30}$$

which leads to

$$\mathbf{S_n} = nf(x)\mathbf{PHSHP}(1 + o_P(1)). \tag{31}$$

In the same manner, we find that

$$\begin{aligned}
S^*_{n,j} &= ES^*_{n_j} + O_P\left(\sqrt{VarS^*_{n,j}}\right) \\
&= nh^{j-1}(\phi^j(x)\sigma^2(x)f(x)\nu_j + o(1)) + O_P\left(\sqrt{o(n^2h^{2j-2})}\right) \\
&= nh^{j-1}\phi^j(x)\sigma^2(x)f(x)\nu_j(1 + o_P(1))
\end{aligned}$$

and thus

$$\mathbf{S^*_n} = \frac{n}{h}f(x)\sigma^2(x)\mathbf{PHS^*HP}(1 + o_P(1)) \tag{32}$$

and finally assertion (15) by plugging (31) and (32) into (24).

II. Asymptotic conditional bias

Finding an asymptotic expression for

$$bias(\hat{\gamma}|\mathbb{X}) = \mathbf{S_n}^{-1}\mathbf{X}^T\mathbf{Wr} \tag{33}$$

still requires to approximate $\mathbf{r} \equiv (r_i)_{1\le i\le n}$. For all data points within the kernel support we obtain

$$\begin{aligned}
r_i &= m(X_i) - \sum_{j=0}^{p}\gamma_j(\Phi(X_i) - \Phi(x))^j \\
&= \frac{\Psi_{(p+1)}(\zeta_i)}{(p+1)!}(\Phi(X_i) - \Phi(x))^{p+1} \\
&= \gamma_{p+1}(x)(\Phi(X_i) - \Phi(x))^{p+1} + o_P(1)\frac{(\Phi(X_i) - \Phi(x))^{p+1}}{(p+1)!}
\end{aligned}$$

where $\zeta_i \in (X_i, x)$ resp. (x, X_i) exists according to Theorem 1. Finally we calculate

$$\begin{aligned}
bias(\hat{\gamma}|\mathbb{X}) &= \mathbf{S_n}^{-1}\mathbf{X}^T\mathbf{W}\left[(\Phi(X_i) - \Phi(x))^{p+1}(\gamma_{p+1} + o_P(1))\right]_{1\le i\le n} \\
&= \mathbf{S_n}^{-1}\mathbf{c_n}(\gamma_{p+1} + o_P(1)) \\
&= \mathbf{P}^{-1}\mathbf{H}^{-1}\mathbf{S}^{-1}\mathbf{H}^{-1}\mathbf{P}^{-1}\frac{1}{nf(x)}\left\{\gamma_{p+1}\mathbf{c_n} + \begin{pmatrix} o(nh^{p+1}) \\ \vdots \\ o(nh^{2p+1}) \end{pmatrix}\right\}(1 + o_P(1)) \\
&= \mathbf{P}^{-1}\mathbf{H}^{-1}\mathbf{S}^{-1}h^{p+1}\phi^{p+1}(x)\gamma_{p+1}\mathbf{c_p}(1 + o_P(1)),
\end{aligned}$$

by substituting the asymptotic expressions for $S_{n,j}$ (30) in $\mathbf{c_n} := (S_{n,p+1}, \dots, S_{n,2p+1})^T$, and thus (16) is proofed.

Now we proceed to the derivation of b_n which requires to take along some extra terms resulting from higher order expansions. With $(a+hb)^j = a^j + h(ja^{j-1}b + o(1))$ we find that

$$\begin{aligned} ES_{n,j} &= nh^j \int K(u)u^j \left(\phi(x) + \frac{hu}{2}\phi'(\zeta_u)\right)^j (f(x) + huf'(\xi_u))du \\ &= nh^j \int K(u)u^j \left[\phi^j(x) + h\left(\frac{j}{2}\phi^{j-1}(x)u\phi'(\zeta_u) + o(1)\right)\right](f(x) + huf'(\xi_u))du \\ &= nh^j \left[f(x)\phi^j(x)\mu_j + h\left(f'(x)\phi^j(x) + \frac{f(x)}{2}j\phi^{j-1}(x)\phi'(x)\right)\mu_{j+1} + o(h)\right] \end{aligned} \tag{34}$$

with ζ_u and ξ_u according to Taylor's theorem. Plugging (34) and (27) into (29) yields

$$S_{n,j} = nh^j\phi^j(x)\left[f(x)\mu_j + h\left(f'(x) + \frac{f(x)}{2}\frac{\phi'(x)}{\phi(x)}j\right)\mu_{j+1} + o_n\right], \tag{35}$$

where $o_n = o_P(h) + O_P\left(\frac{1}{nh}\right)$, and further

$$\mathbf{S_n} = n\mathbf{PH}\left(f(x)\mathbf{S} + hf'(x)\tilde{\mathbf{S}} + h\frac{f(x)}{2}\frac{\phi'(x)}{\phi(x)}\bar{\mathbf{S}} + o_n\right)\mathbf{HP}. \tag{36}$$

The next task is to derive a higer order expansion for $\mathbf{r}$. With Theorem 1 we obtain

$$\begin{aligned} r_i &= \frac{\Psi_{(p+1)}(x)}{(p+1)!}(\Phi(X_i) - \Phi(x))^{p+1} + \frac{\Psi_{(p+2)}(\zeta_i)}{(p+2)!}(\Phi(X_i) - \Phi(x))^{p+2} \\ &= \gamma_{p+1}(\Phi(X_i) - \Phi(x))^{p+1} + \gamma_{p+2}(\Phi(X_i) - \Phi(x))^{p+2} + \\ &\quad +(\Psi_{(p+2)}(\zeta_i) - \Psi_{(p+2)}(x))\frac{(\Phi(X_i) - \Phi(x))^{p+2}}{(p+2)!} \\ &= (\Phi(X_i) - \Phi(x))^{p+1}\gamma_{p+1} + (\Phi(X_i) - \Phi(x))^{p+2}(\gamma_{p+2} + o_P(1)) \end{aligned}$$

with $\zeta_i \in (X_i, x)$ resp. (x, X_i). Plugging this and (36) into (33) and denoting

$$\mathbf{T_n} := f(x)\mathbf{S} + hf'(x)\tilde{\mathbf{S}} + \frac{f(x)}{2}\frac{\phi'(x)}{\phi(x)}\bar{\mathbf{S}} + o_n$$

leads to

$$
\begin{aligned}
bias(\hat{\gamma}|\mathbb{X}) &= \left[n\mathbf{PHT_nHP}\right]^{-1}\left[\mathbf{c_n}\gamma_{p+1} + \tilde{\mathbf{c}}_\mathbf{n}(\gamma_{p+2} + o_P(1))\right] \\
&= \mathbf{P}^{-1}\mathbf{H}^{-1}\mathbf{T_n}^{-1}h^{p+1}\phi^{p+1}(x)\Bigg[\gamma_{p+1}f(x)\mathbf{c_p} + \\
&\quad + h(\gamma_{p+1}f'(x) + \gamma_{p+2}\phi(x)f(x))\tilde{\mathbf{c}}_\mathbf{p} + \\
&\quad + h\gamma_{p+1}f(x)\frac{\phi'(x)}{2\phi(x)}\bar{\mathbf{c}}_\mathbf{p} + o_n\Bigg], \qquad (37)
\end{aligned}
$$

where the asymptotic expressions (35) are substituted in $\mathbf{c_n}$ and $\tilde{\mathbf{c}}_\mathbf{n} = (S_{n,p+2}, \ldots, S_{n,2p+2})^T$. The matrix $\mathbf{T_n}$ still has to be inverted. Applying the formula

$$
(\mathbf{A} + h\mathbf{B})^{-1} = \mathbf{A}^{-1} - h\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1} + O(h^2).
$$

(for the proof see Lemma 23.2 in Hirzebruch & Scharlau (1991)) yields

$$
\mathbf{T_n}^{-1} = \frac{1}{f(x)}\mathbf{S}^{-1} - h\frac{1}{f(x)}\mathbf{S}^{-1}\left(\frac{f'(x)}{f(x)}\tilde{\mathbf{S}} - \frac{\phi'(x)}{2\phi(x)}\bar{\mathbf{S}}\right)\mathbf{S}^{-1} + o_n, \qquad (38)
$$

and we obtain finally

$$
\begin{aligned}
bias(\hat{\gamma}|\mathbb{X}) &= h^{p+1}\phi^{p+1}(x)\mathbf{P}^{-1}\mathbf{H}^{-1}\Bigg\{\gamma_{p+1}\mathbf{S}^{-1}\mathbf{c_p} + \\
&\quad + h\Bigg[\left(\gamma_{p+1}\frac{f'(x)}{f(x)} + \gamma_{p+2}\phi(x)\right)\mathbf{S}^{-1}\tilde{\mathbf{c}}_\mathbf{p} + \gamma_{p+1}\frac{\phi'(x)}{2\phi(x)}\mathbf{S}^{-1}\bar{\mathbf{c}}_\mathbf{p} + \\
&\quad + \gamma_{p+1}\mathbf{S}^{-1}\left(\frac{f'(x)}{f(x)}\tilde{\mathbf{S}} - \frac{\phi'(x)}{2\phi(x)}\bar{\mathbf{S}}\right)\mathbf{S}^{-1}\mathbf{c_p}\Bigg] + o_n\Bigg\}.
\end{aligned}
$$

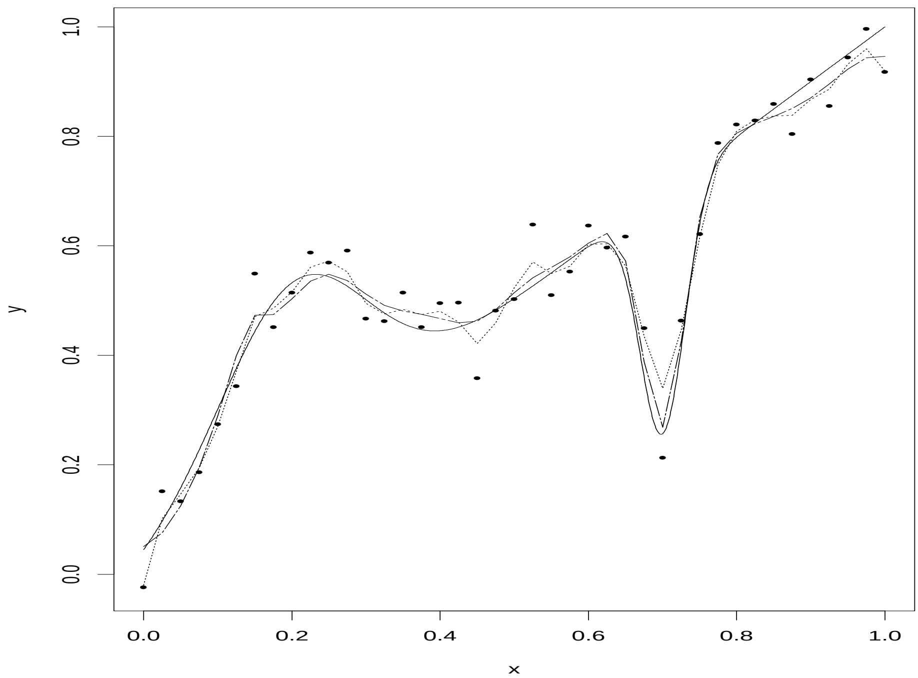


Figure 1: Function (20) (solid line), contaminated data and estimated curve with $\Phi(x) = x$ (dotted line) and $\Phi(x) = \tilde{m}(x)$, (dashed line, see Section 5).

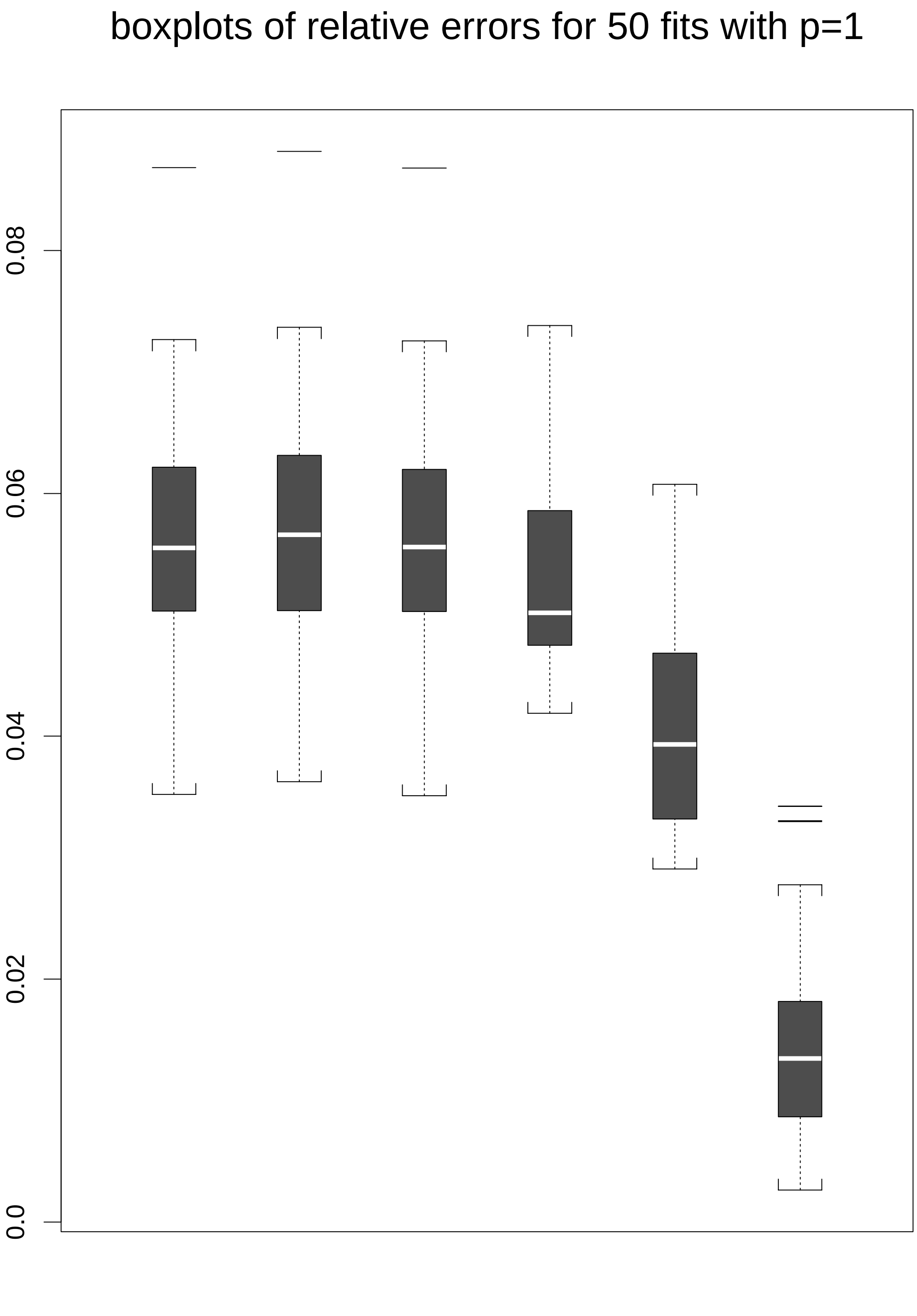


Figure 2: Boxplots of the relative errors using the basis functions $\Phi(x) = x, \cosh(x), \arctan(x), \check{m}(x), \tilde{m}(x), m(x)$ (from left to right) with $p = 1$.

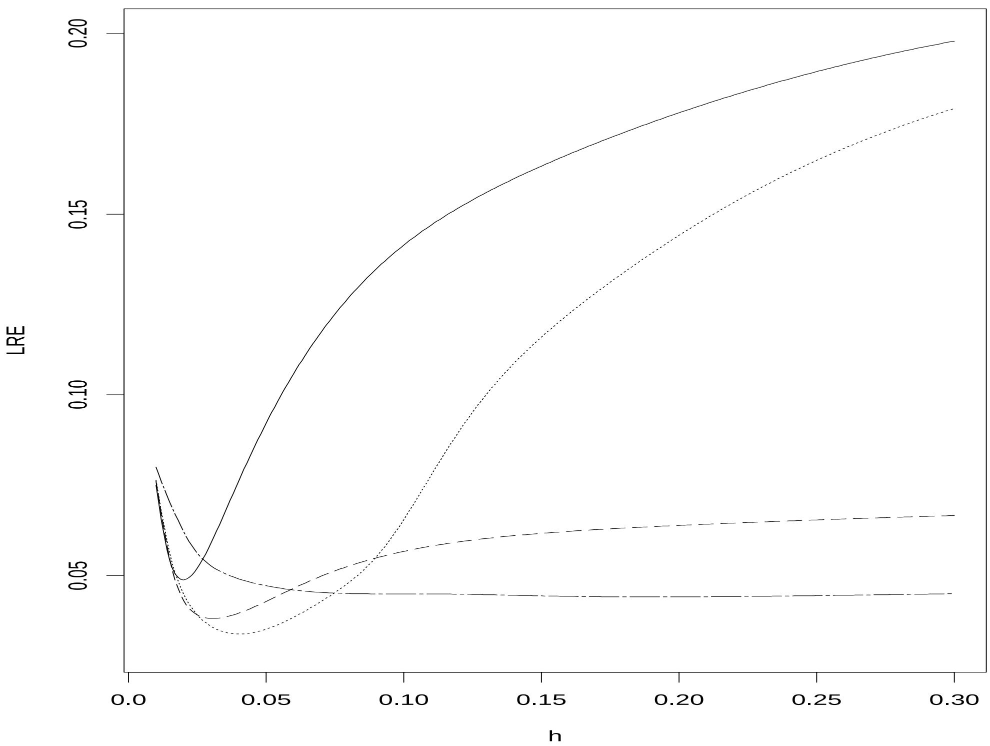


Figure 3: L^2 relative errors as function of the bandwidth for the linear basis (solid line) , the data-adaptive basis $\check{m}_{20}(x)$ (dashed-dotted line) resp. $\check{m}_{38}(x)$ (dashed line), and for the guessed basis $\tilde{m}(x)$ (dotted line).

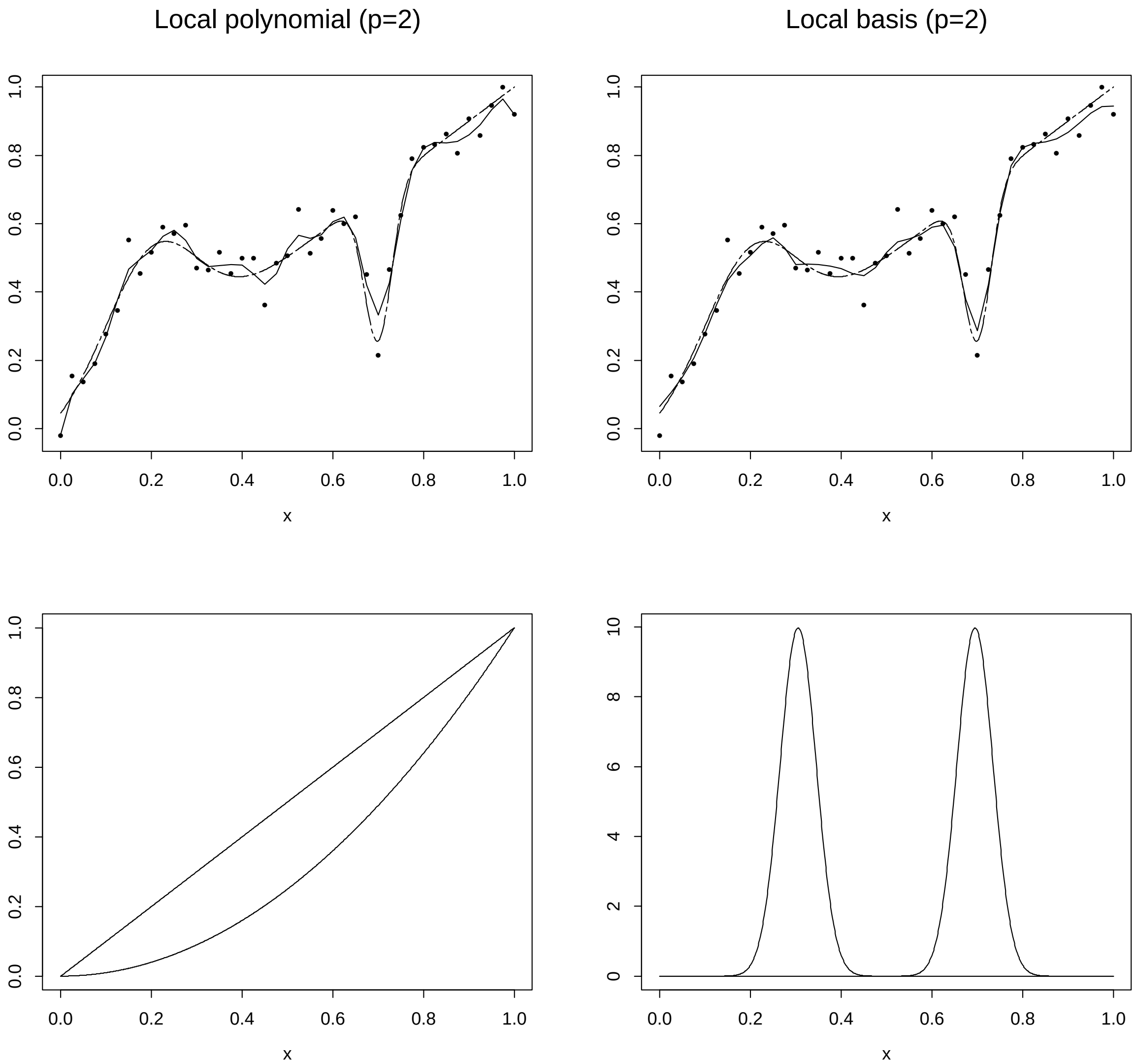
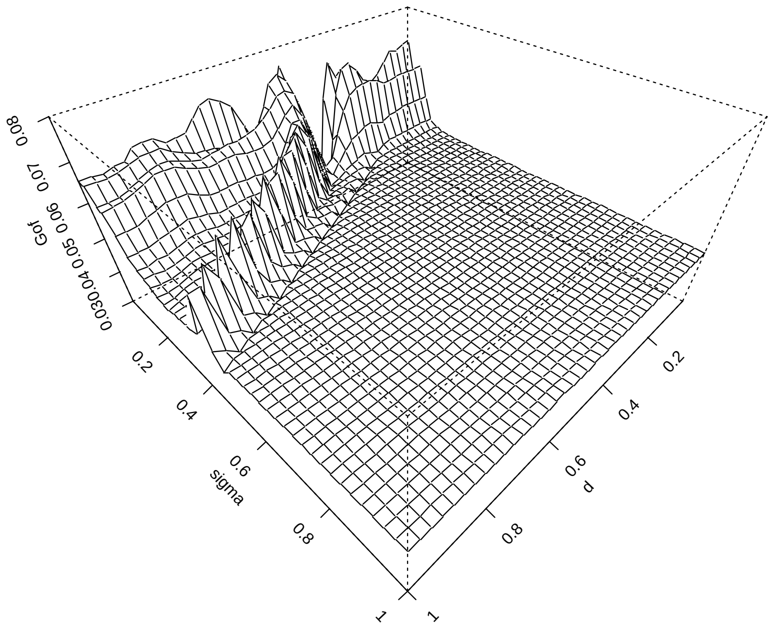


Figure 4: top: data, underlying function (dashed lines) and estimated functions (solid lines) for local polynomial (left) and fitting with a Gaussian basis (right); bottom: basis functions $\Phi_{1,2}(x)$ used for the corresponding fit.



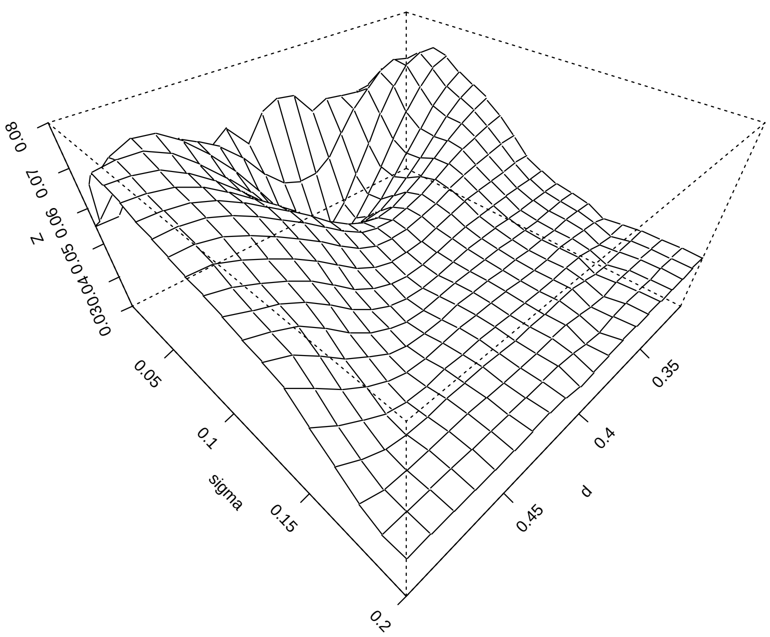


Figure 5: top: L^2 relative errors as function of d and σ (for the function of Example 4 with $h = 0.033$). In the bottom the interesting part in the area of the crater is shown in more depth.

References

Carroll, R. J., Maca, J. D., and Ruppert, D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika* **86**, 541–554.

Choi, E. and Hall, P. (1998). On bias reduction in local linear smoothing. *Biometrika* **85**, 333–345.

Doksum, K., Petersen, D., and Samarov, A. (2000). On variable bandwidth selection in local polynomial regression. *Journal of the Royal Statistical Society, Series B* **62**, 431–448.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications.* London: Chapman & Hall.

Fan, J., Gijbels, I., Hu, T.-C., and Huang, L.-S. (1996). A study of variable bandwidth selection for local polynomial regression. *Statistica Sinica* **6**, 113–127.

Forster, O. (1999). *Analysis 1.* Braunschweig/Wiesbaden: Vieweg.

Hirzebruch, F. and Scharlau, W. (1991). *Einführung in die Funktionalanalysis.* Mannheim: B.I. Wissenschaftsverlag.

Honda, T. (2000). Nonparametric estimation of a conditional quantile for α-mixing processes. *Ann. Inst. Statist. Math.* **52**, 459–470.

Hurvich, C., Simonoff, J., and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B* **60**, 271–293.

Lin, X. and Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor in measured with/without error. *J. Amer. Statist. Assoc.* **95**, 520–534.

Ruppert, D., Sheather, S., and Wand, M. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **90**, 1257–1270.

Seifert, B. and Gasser, T. (2000). Data adaptive ridging in local polynomial regression. *Journal of Comput. and Graph. Statistics* **9**, 338–360.

Yu, K. and Jones, M. (1998). Local linear quantile regression. *J. Amer. Statist. Assoc.* **93**, 228–237.

Zhao, L. (1999). Improved estimators in nonparametric regression problems. *J. Amer. Statist. Assoc.* **94**, 164–173.