REVSTAT – Statistical Journal
Volume 2, Number 2, November 2004

# LOCAL FITTING WITH A POWER BASIS

Authors: Jochen Einbeck
– Department of Statistics, Ludwig Maximilians University Munich, Germany (einbeck@stat.uni-muenchen.de)

Received: April 2003 Revised: September 2004 Accepted: October 2004

Abstract:

- Local polynomial modelling can be seen as a local fit of the data against a polynomial basis. In this paper we extend this method to the power basis, i.e. a basis which consists of the powers of an arbitrary function. Using an extended Taylor theorem, we derive asymptotic expressions for bias and variance of this estimator. We apply this method to a simulated data set for various basis functions and discuss situations where the fit can be improved by using a suitable basis. Finally, some remarks about bandwidth selection are given and the method is applied to real data.

Key-Words:

- *local polynomial fitting; Taylor expansion; power basis; bias reduction.*

AMS Subject Classification:

- 62G08, 62G20.

## 1. INTRODUCTION

The roots of local polynomial modelling as understood today reach back to articles from Stone [19] and Cleveland [1]. A nice overview of the current state of the art is given in Fan & Gijbels [7]. The basic idea of this nonparametric smoothing technique is simply described. Consider bivariate data $(X_1, Y_1), ..., (X_n, Y_n)$, forming an i.i.d. sample from a population $(X, Y)$. Assume the data to be generated from a model

$$Y = m(X) + \sigma(X)\,\varepsilon\ , \tag{1.1}$$

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

$$m(z) \approx \sum_{j=0}^{p} \frac{m^{(j)}(x)}{j!}\,(z-x)^j \equiv \sum_{j=0}^{p} \beta_j(x)\,(z-x)^j\ , \tag{1.2}$$

given that the $(p+1)^{\text{th}}$ derivative of $m(\cdot)$ in a neighbourhood 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, and $h$ denotes the bandwidth. The task of finding the appropriate bandwidth is the crucial point of local polynomial fitting; see Section 6 for more details. Minimizing

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

leads to the locally weighted least squares regression estimator $\hat\beta(x) = (\hat\beta_0(x), ..., \hat\beta_p(x))^T$ and the corresponding estimators

$$\hat m^{(j)}(x) = j!\,\hat\beta_j(x) \tag{1.3}$$

for $m^{(j)}(x)$, $j = 0, ..., p$. Alternative approaches focussed on estimating the conditional quantiles instead of the mean function (Yu & Jones [21], [22]), where a special case is nonparametric robust regression by local linear medians, applying an $L_1$ norm instead of an $L_2$ norm (Fan, Hu & Truong [8]).

Local polynomial modelling can be interpreted as fitting the data locally against the basis functions 1, $X - x$, $(X - x)^2$, ..., $(X - x)^p$. An obviously arising question is now: why should just these basis functions be the best possible ones? In a general framework one may use the basis functions $\phi_0(X)$, $\phi_1(X)$, ..., $\phi_p(X)$, with arbitrary functions $\phi_j : \mathbb{R} \mapsto \mathbb{R}$, $j = 0, ..., p$. However, theoretical results are only available under some restrictions on the basis functions. Regarding (1.2) and (1.3), it is seen that estimation and interpretation of parameters is based on Taylor's expansion. Furthermore, nearly all asymptotic results, e.g. the bias of the estimator, are based on Taylor's theorem. Asymptotics provide a very important tool to find bandwidth selection rules etc., so they play an important role for the use of the estimator in practice.

Thus, if some theoretical background is desired, one needs to develop a new Taylor expansion for every basis one wants 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, analyze the results, and show that the asymptotic bias may be reduced with a suitable choice of the basis. In Section 4 we apply this method to a simulated data set and compare the results for various basis functions. In Section 5 we give some remarks on bandwidth selection. We apply the method on a real data set in Section 6, and finish with a short discussion in Section 7.

## 2. THE POWER BASIS

The family of basis functions that we will treat in this paper is motivated by the following theorem:

**Theorem 2.1** (Taylor expansion for a power basis). *Let $I$ be a non-trivial interval, $m, \phi : I \to \mathbb{R}$, $p{+}1$ times differentiable in $I$, $\phi$ invertible in $I$, and $x \in I$. Then for all $z \in I$ with $z \neq x$, a value $\zeta \in (x,z)$ resp. $(z,x)$ exists such that*

$$m(z) \;=\; \sum_{j=0}^{p} \frac{\psi_{(j)}(x)}{j!} \big(\phi(z)-\phi(x)\big)^j + \frac{\psi_{(p+1)}(\zeta)}{(p+1)!} \big(\phi(z)-\phi(x)\big)^{p+1}$$

*with*

$$\psi_{(j+1)}(\cdot) = \frac{\psi'_{(j)}(\cdot)}{\phi'(\cdot)}\;, \qquad \psi_{(0)}(\cdot) = m(\cdot)\;,$$

*holds.*

The proof is omitted, since this theorem is simply obtained by applying Taylor's theorem, as found for example in Lay ([12], p. 211), on the function $g(\cdot) = (m \circ \phi^{-1})(\cdot)$ at point $\phi(x)$. Assuming the underlying model (1.1), Theorem 2.1 suggests to fit the data locally in a neighborhood of $x$ against the basis functions $1, \phi(X)-\phi(x), ..., (\phi(X)-\phi(x))^p$. We call a basis of this type a *power basis* of order $p$. For $\phi = id$, the power basis reduces to the polynomial basis. For the rest of this paper, we assume that $\phi : \mathbb{R} \to \mathbb{R}$ is $p{+}1$ times differentiable and invertible in a neighborhood of $x$, though the estimation procedure itself, as outlined from (2.5) to (2.7), does not necessarily require this assumption.

Since the parameters

$$\gamma_j(x) := \frac{\psi_{(j)}(x)}{j!}$$

are constructed in a more complex way than the parameters $\beta_j(x)$ for local polynomial fitting, the simple relationship $m^{(j)}(x) = j!\,\beta_j(x)$ cannot be retained. However, by using the simple recursive formula

$$\gamma_j(x) = \frac{1}{j\phi'(x)}\,\gamma'_{j-1}(x)\ , \qquad \gamma_0(x) = m(x)\ ,$$

the parameters $\gamma_j(x)$ ($j \le p$), which we abbreviate by $\gamma_j$ from now on, can be calculated. In this manner the following relations between parameters and the underlying function and their derivatives are derived for the power basis:

$$m(x) = 0!\,\gamma_0 \tag{2.1}$$

$$m'(x) = 1!\,\phi'(x)\,\gamma_1 \tag{2.2}$$

$$m''(x) = 2!\,[\phi'(x)]^2\,\gamma_2 + \phi''(x)\,\gamma_1 \tag{2.3}$$

$$m'''(x) = 3!\,[\phi'(x)]^3\,\gamma_3 + 3!\,\phi''(x)\,\phi'(x)\,\gamma_2 + \phi'''(x)\,\gamma_1 \tag{2.4}$$

$$\vdots$$

Let $w_i(x) = K_h(X_i - x)$. Minimizing

$$\sum_{i=1}^{n}\left\{Y_i - \sum_{j=0}^{p}\gamma_j\big(\phi(X_i) - \phi(x)\big)^j\right\}^2 w_i(x) \tag{2.5}$$

in terms of $(\gamma_0, ..., \gamma_p)$, one obtains the local least squares estimator $\hat{\gamma} = (\hat{\gamma}_0, ..., \hat{\gamma}_p)^T$. The design matrix and the necessary vectors are given by

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

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \gamma = \begin{pmatrix} \gamma_0 \\ \vdots \\ \gamma_p \end{pmatrix} \qquad \text{and} \qquad \mathbf{W} = \begin{pmatrix} w_1(x) & & \\ & \ddots & \\ & & w_n(x) \end{pmatrix}.$$

The minimization problem (2.5) can be written as

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

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 ([7]). Then $\hat{m}(x) = e_1^T\hat{\gamma}$, where $e_1 = (1, 0, ..., 0)^T$, is an estimator for the underlying function $m(\cdot)$ at point $x$. Using (2.2) to (2.4), estimators for the derivatives can be obtained in a similar way. Note that, to ensure that the matrix $\mathbf{X}^T\mathbf{W}\mathbf{X}$ is invertible, at least $p + 1$ design points are required to satisfy $K_h(X_i - x) > 0$. Furthermore it can be shown that

$$Bias(\hat{\gamma}|\mathbb{X}) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{W}\mathbf{r}\ , \tag{2.6}$$

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

$$\text{Var}(\hat{\gamma}|\mathbb{X}) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\,(\mathbf{X}^T\mathbf{\Sigma}\;\mathbf{X})\,(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\ , \tag{2.7}$$

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

## 3. ASYMPTOTICS

Usually formulas (2.6) and (2.7) cannot be used in practice, since they depend on the unknown quantities $\mathbf{r}$ and $\mathbf{\Sigma}$. Consequently an asymptotic derivation is required. We denote

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

for the $j^{\text{th}}$ moments of $K$ and $K^2$. For technical ease we assume that the kernel $K$ is a (not necessarily symmetric) bounded probability density function (i.e. $\mu_0 = 1$) with bounded support, though the latter assumption still can be relaxed significantly (Fan [4], Fan & Gijbels [6]). Further we define the kernel moment matrices

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

Note that the matrix $\mathbf{S}$ is positive definite and thus invertible (Tsybakov [20], Lemma 1). Furthermore we introduce the denotation $\varphi(x) = \phi'(x)$, the matrices $\mathbf{H} = \text{diag}(h^j)_{0\le j\le p}$ and $\mathbf{P} = \text{diag}(\varphi^j(x))_{0\le j\le p}$ and recall that $e_{j+1} = (0, ..., 0, 1, 0, ..., 0)^T$ with 1 at $(j+1)^{\text{th}}$ position. $o_P(1)$ denotes a sequence of random variables which tends to zero in probability, and $O_P(1)$ stands for a sequence of random variables which is bounded in probability. Let $f(\cdot)$ be the design density of $X$. Firstly, we consider interior points, i.e. we assume $x$ to be a fixed point in the support of the design density $f$.

**Theorem 3.1.** *Assume that $f(x) > 0$, $\sigma^2(x) > 0$, $\varphi(x) \neq 0$ 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 $h \to 0$ and $nh \to \infty$. Then the asymptotic conditional covariance matrix of $\hat{\gamma}$ is given by*

$$\text{Var}(\hat{\gamma}|\mathbb{X}) \;=\; \frac{\sigma^2(x)}{nhf(x)}\,\mathbf{P}^{-1}\,\mathbf{H}^{-1}\,\mathbf{S}^{-1}\,\mathbf{S}^*\,\mathbf{S}^{-1}\,\mathbf{H}^{-1}\,\mathbf{P}^{-1}\big(1 + o_P(1)\big)\ . \tag{3.1}$$

*The asymptotic conditional bias is given by*

$$Bias(\hat{\gamma}|\mathbb{X}) \;=\; h^{p+1}\,\varphi^{p+1}(x)\,\mathbf{P}^{-1}\,\mathbf{H}^{-1}\left(\gamma_{p+1}\mathbf{S}^{-1}\mathbf{c_p} + h\mathbf{b}_n\right)\ , \tag{3.2}$$

*where* $\mathbf{b}_n = O_P(1)$*. If in addition* $f'(\cdot)$*,* $m^{(p+2)}(\cdot)$ *and* $\phi^{(p+2)}(\cdot)$ *are continuous in a neighbourhood of x and* $nh^3 \to \infty$*, the sequence* $\mathbf{b}_n$ *can be written as*

$$\mathbf{b}_n = \left(\gamma_{p+1}\frac{f'(x)}{f(x)} + \gamma_{p+2}\,\varphi(x)\right)\mathbf{S}^{-1}\,\tilde{\mathbf{c}}_\mathbf{p} + \gamma_{p+1}\frac{\varphi'(x)}{2\,\varphi(x)}\,\mathbf{S}^{-1}\,\bar{\mathbf{c}}_\mathbf{p} - \gamma_{p+1}\,\mathbf{S}^{-1}\left(\frac{f'(x)}{f(x)}\,\tilde{\mathbf{S}} - \frac{\varphi'(x)}{2\,\varphi(x)}\,\bar{\mathbf{S}}\right)\mathbf{S}^{-1}\,\mathbf{c}_\mathbf{p} + o_P(1)\ . \tag{3.3}$$

This theorem was obtained in the case $\phi(x) = x$ and $p = 1$ by Fan [4], for general $p$ by Ruppert & Wand [16], and for general $\phi(\cdot)$, $p = 1$, and symmetric kernels by Einbeck [3]. Based on Theorem 3.1 and formula (2.1) asymptotic expressions for bias and variance of the estimator of the conditional mean function can be derived. In particular we obtain

$$\begin{aligned}\mathrm{Var}(\hat{m}(x)|\mathbb{X}) &= \mathrm{Var}(e_1^T\hat{\gamma}|\mathbb{X})\\ &= \frac{\sigma^2(x)}{nhf(x)}\,e_1^T\mathbf{S}^{-1}\,\mathbf{S}^*\,\mathbf{S}^{-1}\,e_1\big(1+o_P(1)\big)\ ,\end{aligned} \tag{3.4}$$

which reduces for $p = 1$ to

$$\mathrm{Var}(\hat{m}(x)|\mathbb{X}) = \frac{\sigma^2(x)}{nhf(x)}\,\frac{\displaystyle\int(\mu_2 - u\mu_1)^2\,K^2(u)\,du}{(\mu_0\mu_2-\mu_1^2)^2}\,\big(1+o_P(1)\big)\ . \tag{3.5}$$

It is an important observation that the asymptotic variance of $\hat{m}(\cdot)$ does not depend on the basis function. Next, we take a look at the bias. Using (3.2) and (2.1) one gets

$$\begin{aligned}\mathit{Bias}(\hat{m}(x)|\mathbb{X}) &= \mathit{Bias}(e_1^T\hat{\gamma}|\mathbb{X})\\ &= h^{p+1}\varphi^{p+1}(x)\,e_1^T\left(\frac{\psi_{(p+1)}(x)}{(p+1)!}\,\mathbf{S}^{-1}\,\mathbf{c}_\mathbf{p} + h\mathbf{b}_n\right),\end{aligned} \tag{3.6}$$

reducing for $p = 1$ to

$$\mathit{Bias}(\hat{m}(x)|\mathbb{X}) = \frac{h^2}{2}\left(m''(x) - \frac{\varphi'(x)}{\varphi(x)}\,m'(x)\right)\frac{\mu_2^2-\mu_1\mu_3}{\mu_0\mu_2-\mu_1^2}\,\big(1+o_P(1)\big)\ . \tag{3.7}$$

## 3.1. Derivatives

Similarly, one might take a look at the formulas for the derivatives. Using (2.2), one gets for the derivative estimator for $p = 1$

$$\begin{aligned}\mathrm{Var}\big(\hat{m}'(x)|\mathbb{X}\big) &= \mathrm{Var}\big(\varphi(x)\,e_2^T\hat{\gamma}|\,\mathbb{X}\big)\\ &= \frac{\sigma^2(x)}{nh^3f(x)}\,e_2^T\,\mathbf{S}^{-1}\,\mathbf{S}^*\,\mathbf{S}^{-1}e_2\big(1+o_P(1)\big)\\ &= \frac{\sigma^2(x)}{nh^3f(x)}\,\frac{\displaystyle\int(\mu_1-u\mu_0)^2K^2(u)\,du}{(\mu_0\mu_2-\mu_1^2)^2}\,\big(1+o_P(1)\big)\ ,\end{aligned} \tag{3.8}$$

and

$$
\begin{aligned}
Bias\big(\hat{m}'(x)|\mathbb{X}\big) &= Bias\big(\varphi(x)\, e_2^T \hat{\gamma}|\,\mathbb{X}\big) \\
&= h^p \varphi^{p+1}(x)\, e_2^T \left( \frac{\psi_{(p+1)}(x)}{(p+1)!}\, \mathbf{S}^{-1}\mathbf{c_p} + h\mathbf{b}_n \right) \qquad (3.9)\\
&= \frac{h}{2}\left( m''(x) - \frac{\varphi'(x)}{\varphi(x)}\, m'(x) \right) \frac{\mu_0\mu_3 - \mu_1\mu_2}{\mu_0\mu_2 - \mu_1^2}\, \big(1 + o_P(1)\big) \ ,
\end{aligned}
$$

where (3.8) and (3.9) still hold for general $p$. Looking at (2.3) and (2.4), one might have the impression that the asymptotic formulas for higher derivatives will be extraordinarily complicated. However, first order expansions are easy to derive, since only the leading term $j!\, \varphi^j(x)\, \gamma_j$ determines the asymptotic behaviour. In particular, one gets for arbitrary $j \le p$

$$
\mathrm{Var}\big(\hat{m}^{(j)}(x)|\mathbb{X}\big) = \frac{(j!)^2\, \sigma^2(x)}{nh^{2j+1} f(x)}\, e_{j+1}^T\, \mathbf{S}^{-1}\, \mathbf{S}^*\, \mathbf{S}^{-1}\, e_{j+1}\big(1 + o_P(1)\big)
$$

and

$$
Bias\big(\hat{m}^{(j)}(x)|\mathbb{X}\big) \ = \ h^{p+1-j}\, j!\, \varphi^{p+1}(x)\, e_{j+1}^T \left( \frac{\psi_{(p+1)}(x)}{(p+1)!}\, \mathbf{S}^{-1}\mathbf{c_p} + o_P(1) \right) .
$$

Note that the formula for the variance is identical to the corresponding formula for local polynomial modelling ([7], p. 62), and that the variance is independent of the basis function for any choice of $j$ and $p$.

## 3.2. Design adaption and automatic boundary carpentry

One might wonder why we provided a deeper derivation of $\mathbf{b}_n$ in Theorem 3.1. This is necessary due to a special property of symmetric kernels. Let us consider symmetric kernels throughout the rest of this section. Then, we have $\mu_{2k+1} = \nu_{2k+1} = 0$ for all $k \in \mathbb{N}_0$. The crucial point is that, when estimating the $j^{\text{th}}$ derivative $\hat{m}^{(j)}(\cdot)$, the product $e_{j+1}^T \mathbf{S}^{-1}\mathbf{c_p}$ is zero iff $p - j$ is even. In the case $j = 0$, $p$ even, one gets from (3.6)

$$
Bias(\hat{m}(x)|\mathbb{X}) = \ h^{p+2}\varphi^{p+1}(x)\, e_1^T \mathbf{b}_n \ . \qquad (3.10)
$$

Suppose one increases the order of a power basis from an even order $p$ to an odd order $p+1$. Obviously, the order $O\left(\frac{1}{nh}\right)$ of the variance (3.4) is unaffected, and Fan & Gijbels ([7], p. 77 f) show that the quantity $e_1^T \mathbf{S}^{-1}\mathbf{S}^*\mathbf{S}^{-1} e_1$ remains constant when moving from an even $p$ to $p+1$. Thus, there is not any change in variance. What about the bias? As can be seen from (3.10) and (3.6), the order of the bias remains to be $O(h^{p+2})$. However, for even $p$ the bias involves the design density $f$ and its derivative $f'$, i.e. the estimator is not “design-adaptive” in the sense of Fan [4]. Regarding the case $j = 1$, the situation is similar: the matrix product $e_2^T \mathbf{S}^{-1}\mathbf{S}^*\mathbf{S}^{-1} e_2$ remains constant when moving from an odd $p$ to $p+1$, while the leading term of the bias simplifies. Summarizing, an odd choice of $p-j$ should be preferred to an even choice, and local estimators based on a power basis show exactly the same behavior as local polynomial estimators in terms of design-adaptivity.

Beside this, "odd" local polynomial estimators have another strong advantage compared to "even" ones: they do not suffer from boundary effects and hence do not require boundary corrections. Does this property carry over to estimators based on a power basis as well? We answer this question by considering the case $p=1$ and $j=0$, though the findings remain valid for any odd choice of $p-j$. For a symmetric kernel and an interior point, (3.5) and (3.7) reduce to

$$\text{Var}(\hat{m}(x)|\mathbb{X}) \;=\; \frac{\sigma^2(x)\,\nu_0}{nhf(x)}\big(1+o_P(1)\big) \tag{3.11}$$

and

$$\mathit{Bias}(\hat{m}(x)|\mathbb{X}) \;=\; \frac{h^2\mu_2}{2}\left(m''(x)-\frac{\varphi'(x)}{\varphi(x)}\,m'(x)\right)+o_P(h^2)\;, \tag{3.12}$$

respectively. The variance is exactly the same as for a local linear fit, while the bias expression includes an additional term expressing the interplay between the basis function and the underlying function. Let us consider boundary points now. Without loss of generality we assume that the density $f$ has a bounded support [0;1]. We write a left boundary point as $x=ch$ $(c\geq 0)$, and accordingly a right boundary point as $x=1-ch$. Calculation of the asymptotic bias and variance is straightforward as in Theorem 3.1; the only difference is that kernel moments $\mu_j$ and $\nu_j$ have to be replaced by

$$\mu_{j,c} = \int_{-c}^{\infty} u^j K(u)\,du \qquad \text{and} \qquad \nu_{j,c} = \int_{-c}^{\infty} u^j K^2(u)\,du$$

in case of a left boundary point, and analogously in case of a right boundary point. Thus, the kernel moments never vanish and the problem corresponds to finding bias and variance for asymmetric kernel functions. Indeed, one obtains at $x=ch$

$$\text{Var}(\hat{m}(x)|\mathbb{X}) \;=\; \frac{\sigma^2(0+)}{nhf(0+)}\,\frac{\int(\mu_{2,c}-u\mu_{1,c})^2K^2(u)\,du}{(\mu_{0,c}\,\mu_{2,c}-\mu_{1,c}^2)^2}\big(1+o_P(1)\big) \tag{3.13}$$

and

$$\mathit{Bias}(\hat{m}(x)|\mathbb{X}) = \frac{h^2}{2}\left(m''(0+)-\frac{\varphi'(0+)}{\varphi(0+)}m'(0+)\right)\frac{\mu_{2,c}^2-\mu_{1,c}\,\mu_{3,c}}{\mu_{0,c}\,\mu_{2,c}-\mu_{1,c}^2}\big(1+o_P(1)\big)\,. \tag{3.14}$$

Comparing (3.11) and (3.12) with (3.13) and (3.14) unveils that the rate of the estimator does not depend on the location of the target point $x$. For a nice demonstration of the dependence of the constant factors on $c$ see Fan & Gijbels [5]. For even values of $p-j$, the rate of convergence at boundary points is slower than in the interior.

### 3.3. Bias reduction

According to equation (3.12), the bias of a first-order-fit depends on the basis $\phi(\cdot)$. This effect may be useful for bias reduction. To investigate this, firstly note that (3.12) 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 $\frac{\varphi'(x)}{\varphi(x)}m'(x)$ in (3.12) provides the chance for bias reduction. In the optimal case, the content of the bracket in (3.12) is zero, hence the differential equation

$$m''(x)\,\varphi(x) - m'(x)\,\varphi'(x) \;=\; 0$$

has to be solved, what leads to the solutions

$$\varphi(x) = c_1 m'(x) \qquad (c_1 \in \mathbb{R})$$

and hence

$$\phi(x) = c_1 m(x) + c_2 \qquad (c_1, c_2 \in \mathbb{R})\ . \tag{3.15}$$

Note that for symmetric kernels and $p - j$ odd one has $e_{j+1}^T \mathbf{b}_n = o_P(1)$. Thus, the remaining asymptotic bias is even of order $o_P(h^3)$. Having an optimal basis function in the form of (3.15), one may ask if there is any gain in increasing the order $p$? One finds immediately $\psi_{(1)}(x) = 1/c_1$ and thus

$$\gamma_p(x) = \psi_{(p)}(x)/p! = 0 \qquad \text{for} \;\; p \geq 2\ . \tag{3.16}$$

Thus any additional terms are superfluous, since their parameters should take optimally the value zero. The strategy should consequently be the following: work with $p = 1$, and try to find a basis which is as near as possible to the underlying function.

In particular, for $c_1 = 1$, $c_2 = 0$ we get $\phi_{opt}(x) = m(x)$, thus the underlying function $m(\cdot)$ is a member of the family of optimal basis functions. Certainly, the function $m(\cdot)$ is always unknown. However, there are still at least two ways to use this result. We want to approach them from a philosophical point of view. What does a basis function actually effect? For a given basis, the smoothing step in fact balances between the information given by the basis and the data. A similar concept is well-known from Bayesian statistics (see e.g. Smith & Kohn [18]). Though the Bayesian prior does not contain a basis function but an assumption about the distribution of unknown parameters, the principle, boldly compared, is the same, since the posterior distribution can be interpreted as a trade-off between information in the data and prior knowledge. Thus, having some ("prior") knowledge about $m$, the fitted ("posteriori") curve can be steered in the correct direction when incorporating this knowledge in the basis. If there does not exist any knowledge about $m$, one can calculate a pilot estimate via a local linear fit (or any other smooth fit, e.g. with splines) and use the estimated function as an improved basis. In the following section we will provide examples for the application of these strategies.

## 4. A SIMULATED EXAMPLE

Throughout this section, we consider the underlying function

$$m(x) = x + \frac{1}{1.2\sqrt{2\pi}}\,e^{-(x-0.2)^2/0.02} - \frac{1}{0.9\sqrt{2\pi}}\,e^{-(x-0.7)^2/0.0018}\,, \tag{4.1}$$

which we contaminate with Gaussian noise with $\sigma = 0.05$. The 50 predictors are uniformly distributed on $[0;1]$. We repeated this simulation 50 times, obtaining 50 data sets. See Fig. 3 for getting an impression of the data set. As a measure of performance, we use the relative squared error

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

For each simulated data set and for each estimation $\hat m$ of $m$ with different basis functions and polynomial orders we select the empirically optimal bandwidth $h_{emp}$ by

$$h_{emp} = \min_h \mathrm{RSE}(\hat m)\,.$$

This bandwidth $h_{emp}$ is used for the corresponding fit, and the medians of the 50 RSE values obtained in this manner are shown in Table 1. (Of course, $h_{emp}$ only may be calculated for simulated data. Bandwidth selection for real data is treated in Section 5.) The function dnorm$(x)$ denotes the density of the standard normal distribution. We put a star $(*)$ behind the RSE if the value is better than that for local linear fitting $(\phi(x) = x)$ and two stars for the winner of the column.

**Table 1**: Medians of RSEs for various polynomial orders and basis functions.

| $\phi(x)$ | $p=1$ | $p=2$ | $p=3$ | $p=8$ |
|---|---|---|---|---|
| $x$ | 0.04819 | 0.05005 | 0.04915 | 0.04973 |
| $\sin x$ | 0.04810 $**$ | 0.05003 $*$ | 0.04904 $*$ | 0.05008 |
| $\arctan x$ | 0.04812 $*$ | 0.04997 $*$ | 0.04911 $*$ | 0.05011 |
| $\cosh x$ | 0.04898 | 0.04919 $*$ | 0.04916 | 0.04634 $**$ |
| dnorm $x$ | 0.04893 | 0.04888 $**$ | 0.04844 $**$ | 0.04844 $*$ |
| $\exp x$ | 0.04829 | 0.05005 | 0.04917 | 0.04886 $*$ |
| $\log(x+1)$ | 0.04811 $*$ | 0.04988 $*$ | 0.04917 | 0.05000 |

The corresponding boxplots of the RSE values are presented in Fig. 1. Taking a look at the table and the figure, one notes immediately that the differences between different basis functions are mostly negligible, and the performance does not improve when rising the polynomial order. Looking at the table in more depth, one observes that the group of odd basis functions behaves slightly different than the group of even basis functions. In particular, for $p = 1$ the odd basis functions outperform the even ones. Recalling equation (3.15), this might be interpreted as that the underlying function $m(\cdot)$ possesses rather odd than even characteristics. Finally, one observes that the Gaussian basis yields the best RSE for $p = 2$ and $p = 3$. This is quite intuitive, since the underlying function contains a sum of Gaussians itself.

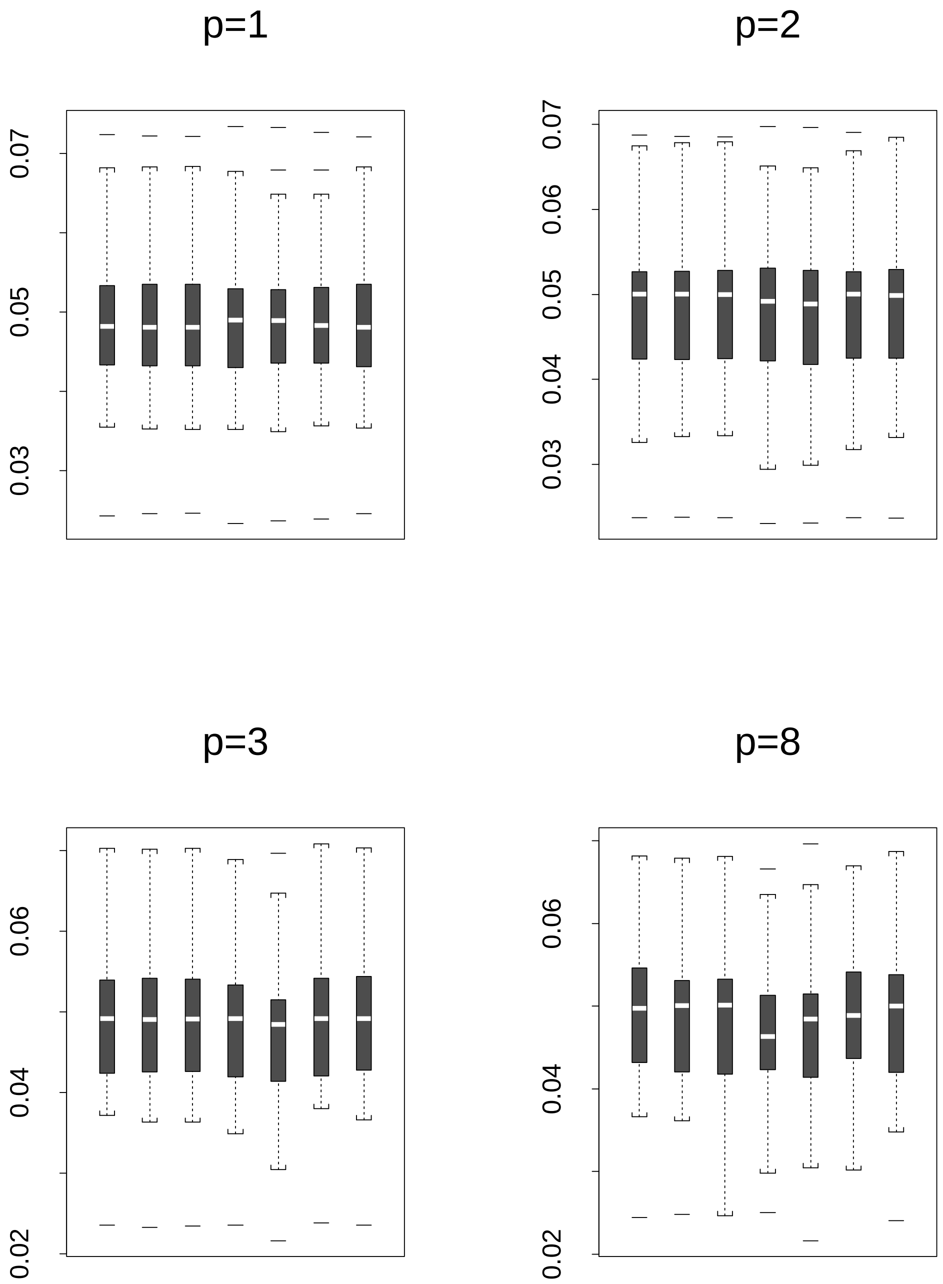

**Figure 1**: Boxplots of the relative errors using the basis functions $\phi(x) = x$, $\sin x$, $\arctan x$, $\cosh x$, dnorm$(x)$, $\exp x$, $\log(x+1)$ and orders $p = 1, 2, 3, 8$.

Next, we will investigate if these results may be improved by the use of basis functions which contain information about the true function, as suggested by (3.15). We distinguish two situations:

a) *Some information about m is available.* We consider exemplarily two cases:

- Assume that the information about the true function is incomplete, e.g. due to a transmission problem, and the true function is only known on the interval $[0.25; 0.75]$ (i.e., only half of the true function is known!). A basis function $m_1(\cdot)$ is constructed by extrapolating the known part of the function by straight lines in a way that the first derivative is continuous.
- Assume somebody gave us a (partly wrong) information about the underlying function (4.1), namely

$$m_2(x) \;=\; x - \frac{1}{1.2\sqrt{2\pi}}\, e^{-(x-0.2)^2/0.02} - \frac{1}{0.9\sqrt{2\pi}}\, e^{-(x-0.7)^2/0.0018} \;,$$

  i.e. the first hump shows down instead of up.

We call basis functions like that “guessed” basis functions.

b) *No information about m is available.* In this case, we employ the pre-fit basis functions $\bar{m}(\cdot)$ and $\check{m}(\cdot)$ calculated with a local constant or linear fit, respectively. Let $g_{emp}$ be the empirically optimal bandwidth of the pre-fit, i.e. $g_{emp} = h_{emp}^{NW}$ for a local constant (*N*adaraya-*W*atson) pre-fit and $g_{emp} = h_{emp}^{LL}$ for a local linear (*LL*) pre-fit. The bandwidth of the pre-fit is then selected as $g = \theta \cdot g_{emp}$, and the second bandwidth as $h = \lambda \cdot h_{emp}^{LL}$, where $\theta$ and $\lambda$ are optimized in terms of RSE on $[1; 2] \times [1; 2]$.

Keeping in mind observation (3.16) and the conclusions drawn from Table 1, we only consider the case $p = 1$ from now on. The medians of 50 RSE values for each basis function are listed in Table 2. For comparison we added the results for the linear basis $\phi(x) = x$ and the (in practice unavailable) optimal basis $\phi(x) = m(x)$. The corresponding boxplots of RSE values are depicted in Fig. 2. In Fig. 3 the basis functions from Table 2 and the corresponding fitted curves are depicted. One notices again: the more similar basis and true function are, the better is the fitted curve. Further, one observes that there is not much gain in using a local linear instead of a local constant pre-fit. The benefit of applying a pre-fit basis is not overwhelming in this example, and is not as impressive as for multivariate predictors ([3]). Taking into account the difficulty of having to select two bandwidths, it is at least questionable if this additional work is worth the effort for univariate predictors. Nevertheless, in the next section we will give some insight in the nature of this two-dimensional bandwidth selection problem.

The “guessed” basis functions lead to a significant improvement, which does not require any extra work compared to a simple local linear fit. This is finally the principal message of this paper: if you have some information, use it in your basis, and your fit will improve. If this basis is wrong, but at least smooth, normally nothing serious should happen, since the commonly applied linear basis is a *wrong* basis as well in the most situations. Replacing one wrong and smooth basis by another wrong and smooth basis function will not make much difference, as demonstrated in Table 1.

**Table 2**: Medians of relative squared errors for improved basis functions.

| $\phi$ | $p = 1$ | |
|---|---|---|
| $x$ | 0.04819 | |
| $\bar{m}(x)$ | 0.04606 | $*$ |
| $\check{m}(x)$ | 0.04538 | $*$ |
| $m_1(x)$ | 0.04488 | $*$ |
| $m_2(x)$ | 0.03758 | $**$ |
| $m(x)$ | 0.01302 | |

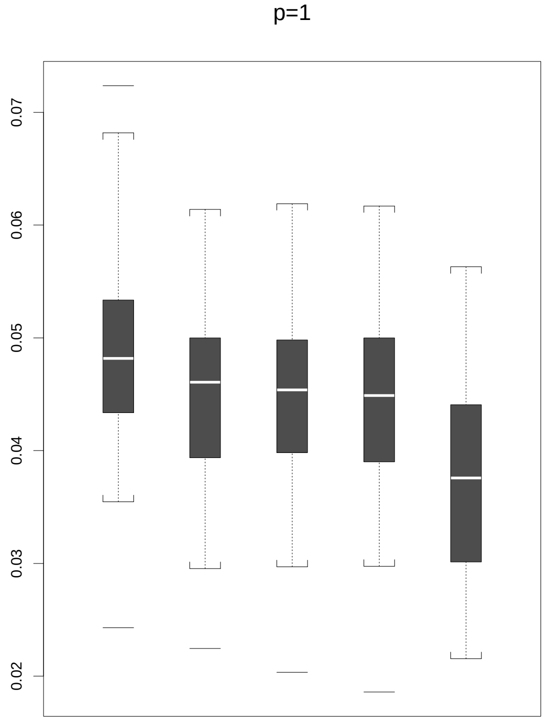

**Figure 2**: Boxplots of the relative errors using the basis functions $\phi(x) = x$, $\bar{m}(x)$, $\check{m}(x)$, $m_1(x)$ and $m_2(x)$ (from left to right) with $p = 1$.

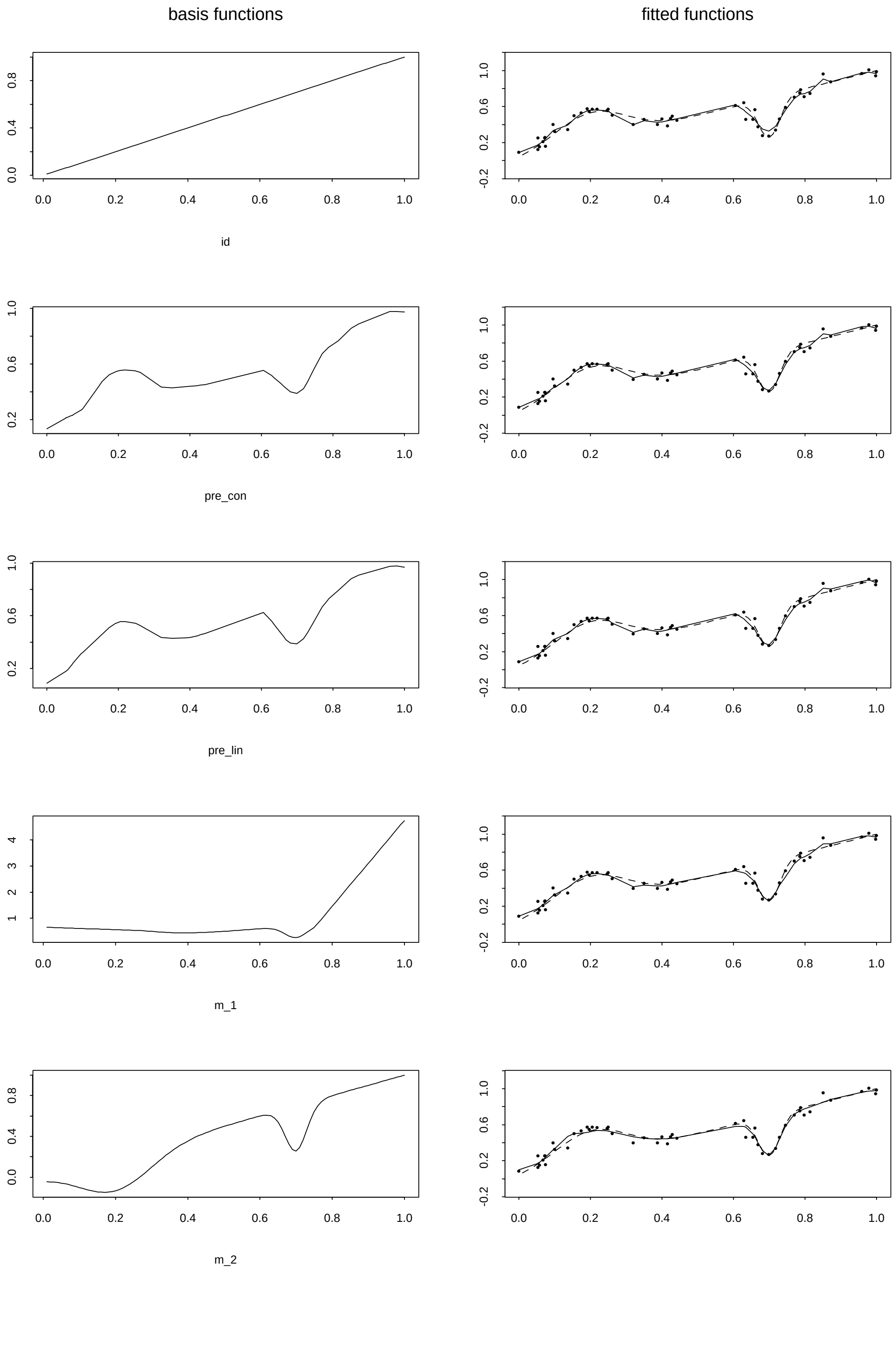

**Figure 3**: Left: basis functions; right: One particular of the 50 simulated data sets $(\cdot)$, true function (dashed line) and fitted functions (solid line) for $p = 1$. The denotations "pre_con" and "pre_lin" refer to the basis functions $\bar{m}$ and $\check{m}$, respectively.

## 5. NOTES ABOUT BANDWIDTH SELECTION

For bandwidth selection, one has the general choice between classical methods and plug-in methods. For an overview of bandwidth selection routines, we refer to Fan & Gijbels ([7], p. 110 ff). Classical methods as cross-validation or the AIC criterion can be applied directly on fitting with general basis functions. Promising extensions of the classical methods have been given by Hart & Yi [9] and Hurvich et al. [11]. In the last decades, classical approaches got the reputation to perform inferior in comparison to plug-in approaches, as treated by Fan & Gijbels [6], Ruppert et al. [15], and Doksum et al. [2], among others. However, this seems not to be justified, as Loader [13] explains, since plug-in-approaches require more theoretical assumptions about the underlying function than classical approaches. Plug-in estimators perform a pilot estimate in order to estimate the asymptotic mean square error, which is then minimized in terms of the bandwidth. Each plug-in-estimator is designed exclusively for a special smoothing method, so that application of these estimators for general basis functions requires some extra work.

Using Theorem 3.1, plug-in formulas for bandwidth selection can be derived straightforwardly by extending the corresponding methods for local polynomial fitting. We will not provide a complete treatment of this topic now, but only give some impressions of the results. Let us therefore consider the derivation of the asymptotically optimal variable bandwidth $h_{opt}(x)$, which varies with the target value $x$. Minimizing the asymptotic mean square error $MSE(\hat{m}(x)|\mathbb{X}) = Bias^2(\hat{m}(x)|\mathbb{X}) + \mathrm{Var}(\hat{m}(x)|\mathbb{X})$ for odd $p-j$, whereby (3.6) and (3.4) are employed for the bias resp. variance, we arrive at an asymptotically optimal bandwidth

$$(5.1) \qquad h_{opt}^{(\phi)}(x) \;=\; C_{0,p}(K) \left[\frac{\sigma^2(x)}{\psi^2_{(p+1)}(x)\, f(x)\, \varphi^{2p+2}(x)}\right]^{\frac{1}{2p+3}} \cdot n^{-\frac{1}{2p+3}} \;,$$

where the constant $C_{0,p}(K)$, which only depends on $p$ and the kernel $K$, is the same as in [7], p. 67. Recall from the end of Section 3 that $\psi_{(p+1)}(x)$ $(p \geq 1)$ approximates zero when $\phi(x)$ approximates $m(x)$. Consequently, the optimal bandwidth tends to infinity when the basis approximates the true function, what is in conformity to the observations which can be drawn from Fig. 4.

Bandwidth selection is especially difficult for data-adaptive basis functions as in the previous section: then we need the two bandwidths $g$ and $h$ for the first and second fit, respectively. We want to give some insight in this bandwidth selection problem, assuming for simplicity that the pre-fit $\bar{m}_g(x)$ is a local constant estimator with constant bandwidth $g$. Intuitively, one would firstly select an (in some sense, e.g. asymptotically) optimal bandwidth $\bar{g}$ of the pre-fit. Afterwards, one would use the resulting fit $\bar{m}_{\bar{g}}$ as a basis for the second fit, applying an optimized bandwidth $h_{opt}^{(\bar{m}_{\bar{g}})}$ for this pre-fit basis. However, this step-wise strategy in practice does not prove to be suitable: when the first fit is too wiggly, the wiggliness carries over to the second fit. Moreover, when the optimal bandwidth is met in the first fit, then the optimal second bandwidth is very high and the minimum of the RSE curve is very flat. In other words: in this case the second fit is superfluous, and the improvement compared to a usual local fit is negligible.

Therefore, it is sensible to use somewhat higher bandwidths in the initial fit. To illustrate this, we return to the example from the previous sections, and examine exemplarily the particular data set depicted in Fig. 3. Following the step-wise strategy outlined above, we select $g = 0.015$ and $h = 0.048$. However, minimizing the RSE simultaneously over $g$ and $h$, one obtains the empirically optimal bandwidth combination (0.030, 0.021). The dependence of the RSE on the bandwidth for different basis functions is demonstrated in Fig. 4. The RSE curve for the initial fit is the solid line, having a minimum at $g = 0.015$ and yielding an estimate $\bar{m}_{15}(x)$. Applying this estimate as a basis function, one gets the dotted line. However, applying the estimate $\bar{m}_{30}(x)$, obtained by a local constant fit with bandwidth $g = 0.030$, one gets the dashed curve. One sees that its minimum is deeper and more localized than that of $\bar{m}_{15}(x)$.

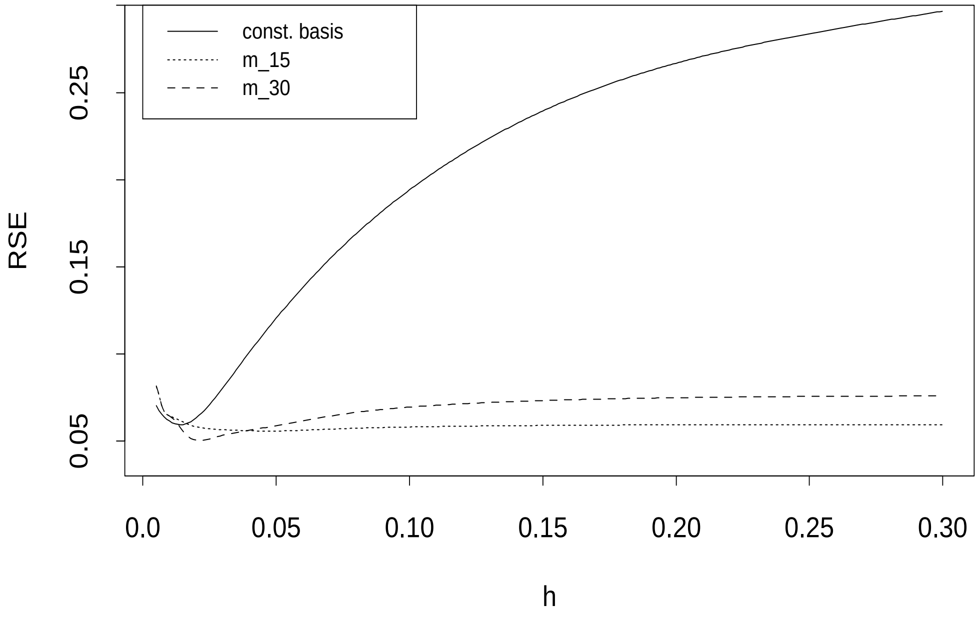

**Figure 4**: RSE as function of the bandwidth for a local constant fit, and for the basis functions $\bar{m}_{15}(x)$, and $\bar{m}_{30}(x)$.

In Section 4 we have already suggested that suitable bandwidths $g$ and $h$ for the pre-fitting algorithm are $1 - 2$ times bigger than the optimal bandwidths of a local constant or a local linear fit, respectively. We want to provide some heuristics to motivate this. Assume that the best bandwidth combination minimizing the RSE simultaneously over $(g, h)$ is given by $(\theta \cdot \bar{g}, \lambda \cdot h_{opt}^{LL})$, where $\bar{g} = h_{opt}^{NW}$. Since we cannot access $\lambda$ directly, we have to apply a sort of trick and to work with a *variable* second bandwidth. Setting (5.1) for $p = 1$ in relation to the optimal variable bandwidth $h_{opt}^{LL}(x)$ for a local linear fit, one obtains

$$\frac{h_{opt}^{(\phi)}(x)}{h_{opt}^{LL}(x)} = \left[1 - \frac{\phi''(x)}{\phi'(x)} \cdot \frac{m'(x)}{m''(x)}\right]^{-2/5} .$$

We define the quantity

$$m^{\circ}(x) = \frac{m''(x)}{m'(x)} ,$$

and substitute for $\phi$ the pre-fit basis $\bar{m}_{\theta\cdot\bar{g}}$. Then one obtains

$$\lambda_x := \frac{h_{opt}^{(\bar{m}_{\theta\cdot\bar{g}})}(x)}{h_{opt}^{LL}(x)} = \left[1 - \frac{\bar{m}^{\circ}_{\theta\cdot\bar{g}}(x)}{m^{\circ}(x)}\right]^{-2/5} \approx \left[1 - \frac{\bar{m}^{\circ}_{\theta\cdot\bar{g}}(x)}{\bar{m}^{\circ}_{\bar{g}}(x)}\right]^{-2/5} . \tag{5.2}$$

What can be said about the relation between $\lambda_x$ and $\theta$? Writing $\bar{m}_g(x) = = \sum_{i=1}^n w_i(x)Y_i / \sum_{i=1}^n w_i(x)$, where $w_i(x) = \frac{1}{g} K\big(\frac{X_i - x}{g}\big)$, one calculates

$$\begin{aligned} \bar{m}^{\circ}_g(x) &= \frac{\bar{m}''_g(x)}{\bar{m}'_g(x)} = \frac{\sum_{i=1}^n w_i{\prime\prime}(x)\big(Y_i - \bar{m}_g(x)\big)}{\sum_{i=1}^n w_i{\prime}(x)\big(Y_i - \bar{m}_g(x)\big)} - 2\,\frac{\sum_{i=1}^n w_i{\prime}(x)}{\sum_{i=1}^n w_i(x)} \\ &= -\frac{1}{g}\,\frac{\sum_{i=1}^n K''\big[(X_i - x)/g\big]\big(Y_i - \bar{m}_g(x)\big)}{\sum_{i=1}^n K'\big[(X_i - x)/g\big]\big(Y_i - \bar{m}_g(x)\big)} + \frac{2}{g}\,\frac{\sum_{i=1}^n K'\big[(X_i - x)/g\big]}{\sum_{i=1}^n K\big[(X_i - x)/g\big]} . \qquad (5.3) \end{aligned}$$

One observes from (5.3) that, roughly approximated,

$$\frac{\bar{m}^{\circ}_{\theta\bar{g}}(x)}{\bar{m}^{\circ}_{\bar{g}}(x)} \approx \frac{1}{\theta} .$$

We substitute this quotient in (5.2) and get

$$\lambda_x \approx \left(1 - \frac{1}{\theta}\right)^{-2/5} \tag{5.4}$$

In order to get a notion about this relation, we assume for a moment equality in (5.4). The function

$$\lambda(\theta) = \big(1 - \theta^{-1}\big)^{-2/5} \tag{5.5}$$

is depicted in Fig. 5 (left). The hyperbolic shape of this function can be observed in reality as well. Let us consider the same data set as utilized in Fig. 4. Performing the pre-fit algorithm for $g, h$ varying on a two-dimensional grid, the resulting RSE values are shown in Fig. 5 (right). The same hyperbola appears again. Thus, the minima of the RSE in terms of the pairs $(g, h)$ are situated along a hyperbola-formed valley. We want to emphasis three special positions in this valley:

- $\theta \to \infty$. Then the first fit is a constant, and the resulting fit is the Nadaraya–Watson-estimator.

- $\lambda \to \infty$. Then the second fit is a parametric regression with a Nadaraya–Watson estimate as basis (which is approximately the same as the previous case).
- $\lambda = \theta$. Then one has $1 = \lambda - \lambda^{-3/2}$, which is solved at about $\lambda = 1.53$. This number corresponds to the magnitude recommended beforehand.

Yet, a generally optimal choice of $\lambda$ and $\theta$ cannot be given. At least we can motivate that the main problem of bandwidth selection for the pre-fitting algorithm can be reduced to the problem of selecting the bandwidth of a local constant or a local linear fit, for the solution of which exist a variety of well established methods. The remaining problem is a problem of fine tuning of the parameters $\theta$ and $\lambda$. Though all considerations in this sections were outlined within the framework of a local constant pre-fit, they remain qualitatively the same for a local linear pre-fit. Indeed, there seems to be no observable advantage of a local linear compared to a local constant pre-fit. Since local constant fitting is more simple than local linear fitting, one might prefer local constant; however, it might be simpler to base both bandwidth selection problems on a local linear fit.

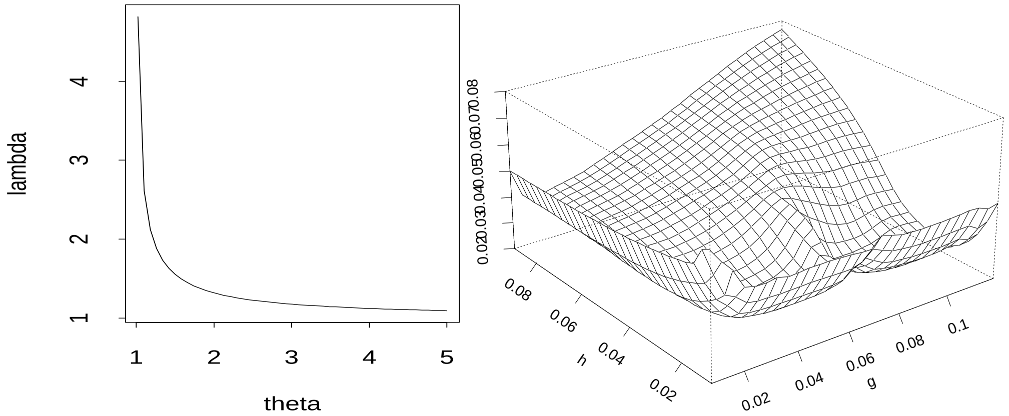

**Figure 5**: Left: function $\lambda(\theta)$; right: RSE for varying $(g, h)$.

## 6. A REAL DATA EXAMPLE

In this section we consider the motorcycle data, firstly provided by Schmidt et al. [17], which have been widely used in the smoothing literature to demonstrate the performance of nonparametric smoothing methods (e.g. [7], p. 2). The data were collected performing crash tests with dummies sitting on motorcycles. The head acceleration of the dummies (in $g$) was recorded a certain time (measured in milliseconds) after they had hit a wall. (Note however that, strictly considered, these data are not fitting the models on which they are usually applied, since there were taken several measurements from every dummy at different time points — thus the data possess an inherent dependence structure. As done in the other literature, we will ignore this problem in the following).

Fig. 6 shows the motorcycle data with a local linear fit (solid line). The bandwidth value 1.48 is obtained by cross-validation. According to the previous sections, the bandwidths $g$ and $h$ should be selected from the interval $[1.48; 2.96]$. Visually, the setting $g = h = 2.6$ was convincing for this data set. The dotted line shows the local linear pre-fit, and the dashed line is the local fit obtained using the pre-fit as basis function. For comparison, we also provide the result of a fit with smoothing splines.

For real data it is hard to judge which fit might be the best one — but at least it seems that the fit applying a local pre-fit basis is less biased at the first bend and the first hump, achieving at the same time a higher smoothness in the outer right area than a local linear fit. The performance seems now comparable to a spline fit.

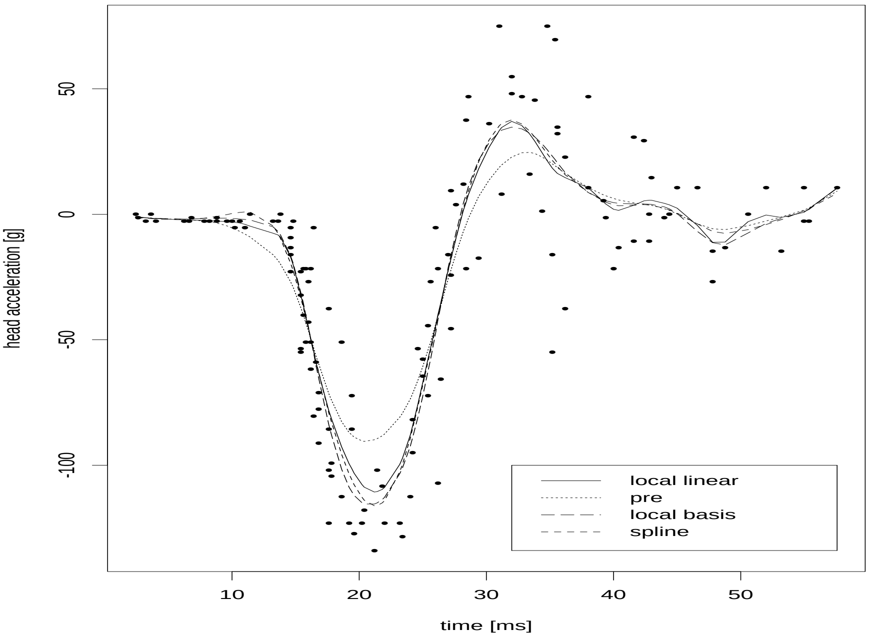

**Figure 6**: Motorcycle data with a local linear fit, a local pre-fit, a local fit using the latter fit as basis function, and a spline fit.

## 7. DISCUSSION

In a certain sense, the main findings of this paper are quite naive. Certainly, everyone has the notion that, when instilling more information about the true function in the basis, the resulting fit should improve. However, it seems that this notion never has been concretized neither from a theoretical nor from a practical point of view, though related ideas have already been mentioned in Ramsay & Silverman ([14], Section 3.3.3) and Hastie & Loader [10]. The main purpose of this work was to fill this gap, and we could confirm the intuitive notion by theoretical as well as practical results. Summarizing, bias reduction is definitely possible when using suitable basis functions, and the possible gain is much bigger than the possible loss by using wrong, but smooth, basis functions. However, application of the pre-fit algorithm can not be unrestrictedly recommended in general, since the possible gain compared to the effort is not overwhelming, at least in the univariate case.

In the framework of this paper it was not possible to solve all open questions completely. There remain some open problems especially concerning bandwidth selection in the case of pre-fitting. Furthermore, it would be useful to know when pre-fitting yields to a significant improvement and when not.

## A. APPENDIX

### Proof of Theorem 3.1

*I. Asymptotic conditional variance*

Whenever there appears in integral in this proof, the borders $-\infty$ and $\infty$ are omitted. We denote $S_{n,j} = \sum_{i=1}^n w_i(x)\,(\phi(X_i) - \phi(x))^j$ and $S^*_{n,j} = \sum_{i=1}^n w_i^2(x)\,\sigma^2(X_i)\,(\phi(X_i) - \phi(x))^j$. 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 (2.7) can be written as

$$\text{(A.1)}\qquad \mathrm{Var}(\hat\gamma|\mathbb{X}) = \mathbf{S_n}^{-1}\,\mathbf{S^*_n}\,\mathbf{S_n}^{-1}$$

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

$$\int K(u)\,u^j\,g(x+hu)\,du \;=\; \mu_j\,g(x) + o(1)$$

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

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

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

$$\begin{aligned}
\operatorname{Var} S_{n,j} &= nE\Big(w_1^2\big(\phi(X_1)-\phi(x)\big)^{2j}\Big) - nE^2\Big(w_1\big(\phi(X_1)-\phi(x)\big)^{j}\Big)\\
&= nh^{2j-1}\Big(f(x)\,\varphi^{2j}(x)\,\nu_{2j} + o(1)\Big)\\
&= n^2h^{2j}O\bigg(\frac{1}{nh}\bigg) \qquad \text{(A.2)}\\
&= o\big(n^2h^{2j}\big)\ .
\end{aligned}$$

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

$$\text{(A.3)} \qquad Y_n = EY_n + O_P\left(\sqrt{\operatorname{Var} Y_n}\right)$$

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

$$\begin{aligned}
S_{n,j} &= ES_{n_j} + O_P\left(\sqrt{\operatorname{Var} S_{n,j}}\right)\\
&= nh^j f(x)\,\varphi^j(x)\,\mu_j(1+o_P(1)) \qquad \text{(A.4)}
\end{aligned}$$

which leads to

$$\text{(A.5)} \qquad \mathbf{S_n} = nf(x)\,\mathbf{P\,H\,S\,H\,P}\,(1+o_P(1))\ .$$

In the same manner, we find that

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

and thus

$$\text{(A.6)} \qquad \mathbf{S^*_n} = \frac{n}{h}\,f(x)\,\sigma^2(x)\,\mathbf{P\,H\,S^*H\,P}\,(1+o_P(1))$$

and finally assertion (3.1) by plugging (A.5) and (A.6) into (A.1).

*II. Asymptotic conditional bias*

Finding an asymptotic expression for

$$\text{(A.7)} \qquad \mathit{Bias}(\hat\gamma|\mathbb{X}) \;=\; \mathbf{S_n}^{-1}\,\mathbf{X}^T\,\mathbf{Wr}$$

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

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

where $\zeta_i \in (X_i, x)$ resp. $(x, X_i)$ exists according to Theorem 2.1, and the term $o_P(1)$ is uniform over $D_K(x)$. Note that the invertibility demanded for $\phi(\cdot)$ in Theorem 2.1 is already guaranteed locally around $x$ by the condition $\varphi(x) \neq 0$. Finally we calculate

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

by substituting the asymptotic expressions for $S_{n,j}$ (A.4) in $\mathbf{c_n} := (S_{n,p+1}, ..., S_{n,2p+1})^T$, and thus (3.2) is proven.

Now we proceed to the derivation of $\mathbf{b}_n$ 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

$$\begin{aligned}
ES_{n,j} &= nh^j \int K(u)\, u^j \left(\varphi(x) + \frac{hu}{2}\, \varphi'(\zeta_u)\right)^j \Big(f(x) + huf'(\xi_u)\Big)\, du \\
&= nh^j \int K(u)\, u^j \left[\varphi^j(x) + h\left(\frac{j}{2}\, \varphi^{j-1}(x)\, u\, \varphi'(\zeta_u) + o(1)\right)\right]\Big(f(x) + huf'(\xi_u)\Big)\, du \\
&= nh^j \left[f(x)\, \varphi^j(x)\, \mu_j + h\Big(f'(x)\, \varphi^j(x) + \frac{f(x)}{2}\, j\, \varphi^{j-1}(x)\, \varphi'(x)\Big)\, \mu_{j+1} + o(h)\right]
\end{aligned} \tag{A.8}$$

with $\zeta_u$ and $\xi_u$ according to Taylor's theorem. Plugging (A.8) and (A.2) into (A.3) yields

$$S_{n,j} \;=\; nh^j\varphi^j(x)\left[f(x)\mu_j + h\left(f'(x) + \frac{f(x)}{2}\,\frac{\varphi'(x)}{\varphi(x)}\, j\right)\mu_{j+1} + o_n\right], \tag{A.9}$$

where $o_n = o_P(h) + O_P\left(\frac{1}{\sqrt{nh}}\right) = o_P(h)$ from the hypothesis $nh^3 \to \infty$, and further

$$\mathbf{S_n} \;=\; n\mathbf{PH}\left(f(x)\mathbf{S} + hf'(x)\tilde{\mathbf{S}} + h\frac{f(x)}{2}\frac{\varphi'(x)}{\varphi(x)}\,\bar{\mathbf{S}} + o_P(h)\right)\mathbf{HP}\ . \tag{A.10}$$

The next task is to derive a higher order expansion for $\mathbf{r}$. With Theorem 2.1 we obtain

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

with $\zeta_i \in (X_i, x)$ resp. $(x, X_i)$. Plugging this and (A.10) into (A.7) and denoting

$$\mathbf{T_n} \;:=\; f(x)\mathbf{S} + h\left(f'(x)\tilde{\mathbf{S}} + \frac{f(x)}{2}\frac{\varphi'(x)}{\varphi(x)}\bar{\mathbf{S}}\right) + o_P(h)$$

leads to

$$\begin{aligned}
\mathit{Bias}(\hat{\gamma}|\mathbb{X}) \;&=\; [n\,\mathbf{PHT_nHP}]^{-1}\Big[\mathbf{c_n}\gamma_{p+1} + \tilde{\mathbf{c}}_\mathbf{n}(\gamma_{p+2} + o_P(1))\Big]\\
&=\; \mathbf{P}^{-1}\mathbf{H}^{-1}\mathbf{T_n}^{-1}h^{p+1}\varphi^{p+1}(x)\,\cdot\\
&\quad\cdot\,\Bigg[\gamma_{p+1}f(x)\,\mathbf{c_p} + h\Big(\gamma_{p+1}f'(x) + \gamma_{p+2}\,\varphi(x)f(x)\Big)\,\tilde{\mathbf{c}}_\mathbf{p}\\
&\qquad + h\,\gamma_{p+1}f(x)\,\frac{\varphi'(x)}{2\,\varphi(x)}\,\bar{\mathbf{c}}_\mathbf{p} + o_P(h)\Bigg]\;,
\end{aligned}$$

where the asymptotic expressions (A.9) are substituted in $\mathbf{c_n}$ and $\tilde{\mathbf{c}}_\mathbf{n} = (S_{n,p+2}, ..., S_{n,2p+2})^T$. The matrix $\mathbf{T_n}$ still has to be inverted. Applying the formula

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

yields

$$\mathbf{T_n}^{-1} \;=\; \frac{1}{f(x)}\mathbf{S}^{-1} - h\frac{1}{f(x)}\mathbf{S}^{-1}\left(\frac{f'(x)}{f(x)}\tilde{\mathbf{S}} - \frac{\varphi'(x)}{2\varphi(x)}\bar{\mathbf{S}}\right)\mathbf{S}^{-1} + o_P(h)\;,$$

and we obtain finally

$$\begin{aligned}
\mathit{Bias}(\hat{\gamma}|\mathbb{X}) &= h^{p+1}\varphi^{p+1}(x)\,\mathbf{P}^{-1}\mathbf{H}^{-1}\,\cdot\\
&\quad\cdot\left\{\gamma_{p+1}\mathbf{S}^{-1}\mathbf{c_p} + h\left[\left(\gamma_{p+1}\frac{f'(x)}{f(x)} + \gamma_{p+2}\,\varphi(x)\right)\mathbf{S}^{-1}\tilde{\mathbf{c}}_\mathbf{p}\right.\right.\\
&\qquad\left.\left. + \;\gamma_{p+1}\,\frac{\varphi'(x)}{2\varphi(x)}\mathbf{S}^{-1}\bar{\mathbf{c}}_\mathbf{p} + \gamma_{p+1}\mathbf{S}^{-1}\left(\frac{f'(x)}{f(x)}\tilde{\mathbf{S}} - \frac{\varphi'(x)}{2\varphi(x)}\bar{\mathbf{S}}\right)\mathbf{S}^{-1}\mathbf{c_p}\right] + o_P(h)\right\}.
\end{aligned}$$

## ACKNOWLEDGMENTS

The author is grateful to Gerhard Tutz, Torsten Scholz and Klaus Hechenbichler, University of Munich, for the fruitful inspirations and discussions during this work, which led to essential impacts on the development of this paper. Moreover, the author thanks two anonymous referees for a lot of helpful comments, which improved the paper substantially. The support by Sonderforschungsbereich 386 (DFG) in various aspects is gratefully acknowledged.

## REFERENCES

[1] Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association*, **74**, 829–836.

[2] 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.

[3] Einbeck, J. (2003). Multivariate local fitting with general basis functions, *Computational Statistics*, **18**, 185–203.

[4] Fan, J. (1992). Design-adaptive nonparametric regression, *Journal of the American Statistical Association*, **87**, 998–1004.

[5] Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers, *Annals of Statistics*, **20**, 2008-2036.

[6] Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaption, *Journal of the Royal Statistical Society, Series B*, **57**, 371–395.

[7] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*, Chapman and Hall, London.

[8] Fan, J.; Hu, T.-C. and Truong, Y.K. (1994). Robust nonparametric function estimation, *Scandinavian Journal of Statistics* **21**, 433–446.

[9] Hart, J.D. and Yi, S. (1998). One-sided cross-validation, *Journal of the American Statistical Association*, **93**, 620–631.

[10] Hastie, T. and Loader, C. (1993). Rejoinder to: “Local regression: Automatic kernel carpentry”, *Statistical Science*, **8**, 139–143.

[11] Hurvich, C.M.; Simonoff, J.S. 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.

[12] Lay, S.R., (1990). *Analysis with an Introduction to Proof*, Prentice Hall, New Jersey.

[13] Loader, C.R. (1999). Bandwidth selection: Classical or plug-in?, *Annals of Statistics*, **27**, 415–438.

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

[15] Ruppert, D.; Sheather, S.J. and Wand, M.P. (1995). An effective bandwidth selector for local least squares regression, *Journal of the American Statistical Association*, **90**, 1257–1270.

[16] Ruppert, D. and Wand, M.P. (1994). Multivariate locally weighted least squares regression, *Annals of Statistics*, **22**, 1346-1370.

[17] Schmidt, G.; Mattern, R. and Schüler, F. (1981). Biochemical investigation to determine physical and traumatological differentiation criteria for the maximum load capacity of head and vertebral column and without protective helmet under effects of impact, Technical Report Project 65, EEC Research Program on Biomechanics of Impacts, University of Heidelberg, Germany.

[18] Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection, *Journal of Econometrics*, **75**, 317–434.

[19] Stone, C.J. (1977). Consistent nonparametric regression, *Annals of Statistics*, **5**, 595–645.

[20] Tsybakov, A.B. (1986). Robust reconstruction of functions by the local approximation approach, *Problems of Information Transmission*, **22**, 133–146.

[21] Yu, K. and Jones, M.C. (1997). A comparison of local constant and local linear regression quantile estimators, *Computational Statistics and Data Analysis*, **25**, 159–166.

[22] Yu, K. and Jones, M.C. (1998). Local linear quantile regression, *Journal of the American Statistical Association*, **93**, 228–237.