Parametrically Guided Non-parametric Regression

INGRID K. GLAD

University of Oslo

ABSTRACT. We present a new approach to regression function estimation in which a non-parametric regression estimator is guided by a parametric pilot estimate with the aim of reducing the bias. New classes of parametrically guided kernel weighted local polynomial estimators are introduced and formulae for asymptotic expectation and variance, hence approximated mean squared error and mean integrated squared error, are derived. It is shown that the new classes of estimators have the very same large sample variance as the estimators in the standard non-parametric setting, while there is substantial room for reducing the bias if the chosen parametric pilot function belongs to a wide neighbourhood around the true regression line. Bias reduction is discussed in light of examples and simulations.

Key words: bias reduction, correction factor, kernel estimators, local linear regression, local polynomial regression, semiparametric regression

1. Introduction

Suppose that n i.i.d. pairs (X_i, Y_i) are observed from a smooth joint density $p(x, y) = f(x)g(y|x)$. We address the regression problem of estimating the conditional mean function $m(x) = E(Y|X = x)$.

A standard solution to this problem is to fit some parametric model $m(x, \beta)$ to the data. This approach provides an excellent estimator of $m(x)$ if the class of parametric functions happens to be correctly chosen, but can otherwise give a completely wrong picture of the underlying regression function. Non-parametric methods, on the other hand, have in general a slower rate of convergence, but need no explicit specification of the form of the regression function. The resulting curve is hence completely determined by the data themselves. This paper proposes a new regression method where a non-parametric estimator is multiplicatively guided by a parametric pilot estimate, hence combining the two approaches in a semiparametric fashion. The idea builds on the simple identity

$$m(x) = m(x, \beta) \cdot \frac{m(x)}{m(x, \beta)},$$

and proceeds using a parametric estimator $m(x, \hat{\beta})$ for the first factor and a non-parametric estimator for the correction factor $r(x) = m(x)/m(x, \hat{\beta})$. Hence, with a suitable non-parametric estimator $\hat{r}(x)$, the new estimator has the form

$$\hat{m}(x) = m(x, \hat{\beta}) \cdot \hat{r}(x).$$

The key point is that if the parametric guide $m(x, \hat{\beta})$ captures some of the features of the shape of $m(x)$, the correction factor $r(x)$ will be less variable than $m(x)$ itself and hence easier to estimate non-parametrically. Formulae for asymptotic expectation and variance for the new class of regression estimators are derived and compared to the standard non-parametric regression setting. Similar ideas are described in Hjort & Glad (1995, 1996), where a new class of guided

kernel density estimators is suggested and demonstrated to have substantial advantages compared to the purely non-parametric kernel density estimator.

The non-parametric method involved in the estimate of the correction function $r(x)$ might in principle be any of the kernel type regression estimators available in the literature. We have chosen to concentrate on the important class of kernel weighted local polynomial regression estimators, in which the estimated regression curve is obtained by fitting locally in every point x a polynomial of degree p. For readers familiar with local polynomials, we anticipate that with a local polynomial inspired correction function $\hat{r}(x, p)$, the proposed class of guided estimators is of the form

$$\hat{m}(x, p) = m(x, \hat{\beta}) \cdot \hat{r}(x, p) = e_1^{\mathrm{T}} (X_x^{\mathrm{T}} Z_x X_x)^{-1} X_x^{\mathrm{T}} Z_x U_x,$$

with notation to be specified later. The class of local polynomial estimators includes the Nadaraya–Watson estimator ($p = 0$) and the local linear regression estimator ($p = 1$), see Stone (1977), Cleveland (1979) and Fan (1992, 1993), Fan & Gijbels (1992), Hastie & Loader (1993) for theoretical results. For local polynomials of higher order, theoretical results are obtained in Ruppert & Wand (1994). Wand & Jones (1995) and Fan & Gijbels (1996) give excellent accounts of the theory of kernel weighted local polynomial modelling. Alternative versions of these estimators, constructed for example to handle boundary problems, as in Gasser & Müller (1979), or to have a fast implementation (Scott, 1992), can be used as well.

The global parametric pilot estimate $m(x, \hat{\beta})$ might be obtained by any parametric technique, ranging from simple linear regression methods to far more complex techniques such as non-linear regression or regression splines with few knots. However, very often even a simple and rough parametric guide is enough to improve the regression estimate compared to the purely non-parametric versions.

Recent results on semiparametric regression methods include the work by Gozalo & Linton (1996) where any parametric function, rather than a polynomial, is fitted locally with kernel weights. An additive, rather than multiplicative, combination of a parametric estimator and the Nadaraya–Watson estimator is studied in Fan & Ullah (1996), obtaining asymptotic properties similar to those obtained here.

Another multiplicative correction factor method has been independently proposed by Jones *et al.* (1995) in a totally non-parametric form. An initial kernel estimator is multiplicatively corrected with a non-parametric correction function in the same fashion as above. This typically leads to reduction of the bias, but at the cost of a somewhat increased variance. The present approach does not suffer from this trade-off between bias and variance; introducing the parametric guide function allows us to achieve bias reduction while the large sample variance remains unchanged.

The paper is organized in the following way: the estimator with a fixed start function $m(x, \beta_0)$ and a local polynomial correction factor is presented and analysed in section 2. In section 3 we introduce an estimated parametric start function $m(x, \hat{\beta})$. The theory needed to study the asymptotic behaviour of this new class of estimators is introduced, and the properties derived. A multidimensional version of the new estimator and formulae for its asymptotic properties are also provided. In section 4 we assert the general conditions for better asymptotic behaviour of the new class of estimators compared to purely non-parametric methods. These conditions are studied in two specific cases, assuming simple parametric guides for underlying regression functions that are not necessarily of this type. A brief simulation study on finite sample comparison is presented in section 5.

This paper does not address the problem of choosing the required smoothing parameter in the non-parametric estimation step. This question arises in connection with all non-parametric

methods and it is not the aim of this paper to contribute to this issue. If automatic procedures are needed, they can be obtained by following the regimes of automatic selection procedures for the traditional estimators, with some additional complications caused by the parametric function.

2. The new estimator with a fixed guide

Let $m(x)$ be the conditional mean function, where $x \in R$, and let $\sigma^2(x) = \mathrm{var}(Y|X = x)$ be the conditional variance.

A powerful non-parametric method for estimating $m(x)$ is the method of kernel weighted local polynomials. To fix notation, let $\tilde{m}(x, p)$ be the local polynomial regression estimator of degree p. This estimator is obtained by locally fitting a polynomial to the observations (X_i, Y_i) using kernel weighted least squares estimation.

Let Y be the vector of responses and define

$$X_x = \begin{bmatrix} 1 & (X_1 - x) & \dots & (X_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (X_n - x) & \dots & (X_n - x)^p \end{bmatrix}$$

and $Z_x = \mathrm{diag}\{K_h(X_1 - x), \dots, K_h(X_n - x)\}$. Here $K(\cdot)$ is a kernel function, and $K_h(t) = h^{-1}K(th^{-1})$, $h = h(n)$ being the smoothing parameter.

Standard weighted least squares theory then gives the local polynomial regression estimator as

$$\tilde{m}(x, p) = e_1^{\mathrm{T}}(X_x^{\mathrm{T}} Z_x X_x)^{-1} X_x^{\mathrm{T}} Z_x Y \tag{1}$$

where the vector e_1 has 1 in the first entry and 0 in the p other ones.

We collect in condition 1 the general regularity assumptions needed for all expansions in this paper.

Condition 1

Let S be a neighbourhood of the point x.

(a) The marginal density $f \in \mathscr{C}^1(S)$, $|f'| < \infty$, and $f \neq 0$ on S.
(b) The conditional variance $\sigma^2 \in \mathscr{C}(S)$, and $\sigma^2 < \infty$ on S.
(c) The regression function $m \in \mathscr{C}^{p+1}(S)$, and $|m^{(p+1)}| < \infty$ on S for p odd, and $m \in \mathscr{C}^{p+2}(S)$, $|m^{(p+2)}| < \infty$ on S for p even.
(d) The kernel K is a bounded symmetric density function with finite fourth order moment.

Furthermore, we need the following definition of a higher order kernel $K_{(p)}$. Let $\sigma_K^i = \int t^i K(t)\, dt$ and let N_p be the $(p+1) \times (p+1)$ matrix having σ_K^{i+j-2} at entry (i, j). Let $M_p(t)$ be like N_p, but with the first column replaced by the vector $(1, t, \dots, t^p)^{\mathrm{T}}$. Then $K_{(p)}(t) = K(t) \cdot |M_p(t)|/|N_p|$. Note that $K_{(0)} = K_{(1)} = K$.

Conditioned on $(X_1, X_2, \dots, X_n)$ and assuming condition 1, the local polynomial estimator (1) has the following properties as $n \to \infty$, $h \to 0$, and $nh \to \infty$:

for p odd,

$$E(\tilde{m}(x, p)|X_1, \dots, X_n) = m(x) + h^{p+1}\left\{\frac{m^{(p+1)}(x)}{(p+1)!}\right\}\left\{\int t^{p+1} K_{(p)}(t)\, dt\right\} + o_P(h^{(p+1)}),$$

for p even,

$$E(\tilde{m}(x,\ p)|X_1,\ \ldots,\ X_n) = m(x) + h^{p+2}\left\{\frac{m^{(p+2)}(x)}{(p+2)!} + \frac{m^{(p+1)}(x)f'(x)}{(p+1)!f(x)}\right\} \times \left\{\int t^{p+2}K_{(p)}(t)\,dt\right\} + o_P(h^{(p+2)}),$$

and in either case

$$\mathrm{var}(\tilde{m}(x,\ p)|X_1,\ \ldots,\ X_n) = (nh)^{-1}\left\{\frac{\sigma^2(x)}{f(x)}\right\}\left\{\int K_{(p)}^2(t)\,dt\right\} + o_P((nh)^{-1}),$$

see Ruppert & Wand (1994).

Now let $m_0(x)$, $x \in R$, be any non-random function meant to roughly approximate $m(x)$. The correction function $r(x) = m(x)/m_0(x)$ can be non-parametrically estimated by the local polynomial estimator

$$\hat{r}(x,\ p) = e_1^{\mathrm{T}}(X_x^{\mathrm{T}}Z_xX_x)^{-1}X_x^{\mathrm{T}}Z_xU,$$

where U is the vector $(Y_1/m_0(X_1),\ \ldots,\ Y_n/m_0(X_n))^{\mathrm{T}}$. This construction is intuitively motivated by the fact that

$$E\left(\frac{Y_i}{m_0(X_i)}|X_i = x_i\right) = \frac{m(x_i)}{m_0(x_i)} = r(x_i).$$

(This is well defined for $m_0 \neq 0$ on R.) Combining this estimated correction function with the fixed start function $m_0(x)$, we propose the guided local polynomial estimator for the conditional mean regression function

$$\hat{m}(x,\ p) = m_0(x)\cdot\hat{r}(x,\ p) = e_1^{\mathrm{T}}(X_x^{\mathrm{T}}Z_xX_x)^{-1}X_x^{\mathrm{T}}Z_xU_x, \tag{2}$$

where

$$U_x = \left(Y_1\frac{m_0(x)}{m_0(X_1)},\ \ldots,\ Y_n\frac{m_0(x)}{m_0(X_n)}\right)^{\mathrm{T}}. \tag{3}$$

To exemplify, the parametrically guided Nadaraya–Watson estimator reads

$$\hat{m}(x,\ 0) = \sum_{i=1}^{n} Y_i\frac{m_0(x)}{m_0(X_i)}K_h(X_i - x)\Big/\sum_{i=1}^{n}K_h(X_i - x),$$

while the guided local linear estimator becomes

$$\hat{m}(x,\ 1) = \frac{1}{n}\sum_{i=1}^{n} Y_i\frac{m_0(x)}{m_0(X_i)}\{s_2(x) - s_1(x)(X_i - x)\}K_h(X_i - x)/(s_2(x)s_0(x) - s_1(x)^2)$$

where

$$s_k(x) = \frac{1}{n}\sum_{i=1}^{n}(X_i - x)^k K_h(X_i - x), \quad k = 0,\ 1,\ 2.$$

Note that these estimators are completely straightforward to implement. The traditional local polynomial estimators are obtained as a special case by choosing $m_0(x) = c$, a constant. Hence the new class of estimators can be considered as a generalization of the usual non-parametric estimators.

Theorem 1

Let $m_0 \in \mathscr{C}^{p+2}(S)$ be some fixed function satisfying $|m_0| > \delta > 0$ on R. Under condition 1, as $n \to \infty$, $h \to 0$, and $nh \to \infty$, the generalized local polynomial estimator $\hat{m}(x, p)$ in (2) satisfies

for p odd,

$$E(\hat{m}(x, p)|X_1, \ldots, X_n) = m(x) + h^{p+1}\left\{\frac{m_0(x)r^{(p+1)}(x)}{(p+1)!}\right\}\left\{\int t^{p+1}K_{(p)}(t)\,dt\right\} + o_P(h^{(p+1)}),$$

for p even,

$$E(\hat{m}(x, p)|X_1, \ldots, X_n) = m(x) + h^{p+2}\left\{\frac{m_0(x)r^{(p+2)}(x)}{(p+2)!} + \frac{m_0(x)r^{(p+1)}(x)f'(x)}{(p+1)!f(x)}\right\} \times \left\{\int t^{p+2}K_{(p)}(t)\,dt\right\} + o_P(h^{(p+2)}),$$

and in either case

$$\mathrm{var}(\hat{m}(x, p)|X_1, \ldots, X_n) = (nh)^{-1}\left\{\frac{\sigma^2(x)}{f(x)}\right\}\left\{\int K_{(p)}^2(t)\,dt\right\} + o_P((nh)^{-1}).$$

Proof. We show here the proof for $p = 1$. The proof for general p follows the same outline, but becomes less clear due to the more intricate notation. See Ruppert & Wand (1994).

Starting with the expectation, $\hat{m}(x, 1)$ as in (2) has conditional expectation

$$E(\hat{m}(x, 1)|X_1, \ldots, X_n) = m_0(x)e_1^{\mathrm{T}}(X_x^{\mathrm{T}}Z_xX_x)^{-1}X_x^{\mathrm{T}}Z_xR$$

where $R = (r(X_1), \ldots, r(X_n))^{\mathrm{T}}$.

Following the spirit of the conditional proof for the traditional local linear estimator as presented in Wand & Jones (1995), but being slightly more explicit, we expand the elements of R around x to obtain

$$R = X_x\begin{bmatrix} r(x) \\ r'(x) \end{bmatrix} + \tfrac{1}{2}r''(x)\begin{bmatrix} (X_1 - x)^2 \\ \vdots \\ (X_n - x)^2 \end{bmatrix} + \begin{bmatrix} o_P((X_1 - x)^2) \\ \vdots \\ o_P((X_n - x)^2) \end{bmatrix},$$

using Young's form of Taylor's theorem, see for example Serfling (1980, sect. 1.12) and since the diagonal elements of Z_x are asymptotically non-zero only when $X_i - x = O_P(h)$. Insertion of the expansion above then gives

$$E(\hat{m}(x, 1)|X_1, \ldots, X_n) = m_0(x)r(x) + \tfrac{1}{2}m_0(x)r''(x)e_1^{\mathrm{T}}(X_x^{\mathrm{T}}Z_xX_x)^{-1}X_x^{\mathrm{T}}Z_x\begin{bmatrix} (X_1 - x)^2 \\ \vdots \\ (X_n - x)^2 \end{bmatrix} + o_P(h^2).$$

The matrix expression to the right is exactly equal to that appearing in the proof of Wand

& Jones (1995, sect. 5.3.2) (adapted from Fan, 1992); hence by the same arguments we have proved that

$$E(\hat{m}(x, 1)|X_1, \ldots, X_n) = m(x) + h^2\left\{\frac{m_0(x)r''(x)}{2}\right\}\left\{\int t^2 K(t)\,dt\right\} + o_P(h^2).$$

Turning to the variance, we have that

$$\mathrm{var}(\hat{m}(x, 1)|X_1, \ldots, X_n) = e_1^{\mathrm{T}}(X_x^{\mathrm{T}} Z_x X_x)^{-1} X_x^{\mathrm{T}} Z_x C_x Z_x X_x (X_x^{\mathrm{T}} Z_x X_x)^{-1} e_1$$

where C_x is the $(n \times n)$ conditional variance–covariance matrix of U_x with entries $\{c_x(X_i, X_j)\}$. Furthermore,

$$n^{-2} X_x^{\mathrm{T}} Z_x C_x Z_x X_x = n^{-2} \sum_{i=1}^{n}\sum_{j=1}^{n} c_x(X_i, X_j) K_h(X_i - x) K_h(X_j - x) \cdot \begin{bmatrix} 1 & (X_j - x) \\ (X_i - x) & (X_i - x)(X_j - x) \end{bmatrix}, \tag{4}$$

while the expression for $(X_x^{\mathrm{T}} Z_x X_x)^{-1}$ is given in Wand & Jones (1995, p124). But

$$C_x = \mathrm{diag}\left\{\frac{m_0^2(x)}{m_0^2(X_1)}\sigma^2(X_1), \ldots, \frac{m_0^2(x)}{m_0^2(X_n)}\sigma^2(X_n)\right\},$$

hence $c_x(X_i, X_j) = 0$ for all $i \neq j$ and (4) simplifies considerably. By writing $\sigma^2(X_i)m_0^2(x)/m_0^2(X_i) = v(X_i)$ for every i and approximating the averages of the right hand side of (4) with expectations (using Taylor expansions), we obtain

$$\mathrm{var}(\hat{m}(x, 1)|X_1, \ldots, X_n) = (nh)^{-1}\left\{\frac{v(x)}{f(x)}\right\}\left\{\int K^2(t)\,dt\right\} + o_P((nh)^{-1})$$

through the needed matrix multiplications. Finally, since $v(x) = \sigma^2(x)$, we get that $\mathrm{var}(\hat{m}(x, 1)|X_1, \ldots, X_n)$ is equal to the variance stated in theorem 1.

Comparing with the same expressions for the standard local polynomial estimator, we see that for the same h and K, the asymptotic variance remains the very same up to the order used, while there is room for bias reduction if one can choose $m_0(x)$ in such a way that

$$|m_0(x)r^{(p+1)}(x)| < |m^{(p+1)}(x)| \tag{5}$$

for odd p, or

$$\left|m_0(x)r^{(p+2)}(x) + (p+2)m_0(x)r^{(p+1)}\frac{f'(x)}{f(x)}\right| < \left|m^{(p+2)}(x) + (p+2)m^{(p+1)}\frac{f'(x)}{f(x)}\right| \tag{6}$$

for even p. If the initial choice $m_0(x)$ happens to be proportional to the true one, $m_0(x) = c \cdot m(x)$, the correction factor $r(x)$ is constant and hence the bias reduces to a smaller order. If $m_0(x)$ captures some of the features of or is close to $m(x)$ (in senses to be discussed later), $r(x)$ will be less rough than $m(x)$ itself, causing bias reduction while keeping the same asymptotic variance as the traditional non-parametric estimators. We return to such comparisons in later sections.

Remark 1. The positivity assumption (in absolute value) on $m_0(x)$ is typical in regression contexts, as mentioned in Jones *et al.* (1995). Unlike density functions, regression functions $m_0(x)$ might of course in practice cross the x-axis one or more times, causing local problems for

the guided estimator above. If some $m_0(X_i)$ is zero or almost zero while $m_0(x)$ is not, and X_i is close enough to x to receive significant weight from the kernel, the fraction $m_0(x)/m_0(X_i)$ necessarily blows up. This happens in some interval (depending on h) around the point of intersection with the axis. Additionally, in such areas the fractions might obtain negative signs when x is on one side of the point of intersection while X_i is on the other. These effects are only local, however; away from such intervals the estimator works perfectly well and in accordance with the stated theory. An effective way of making the estimator more robust is to use the absolute value together with a suitable truncation of the fraction, for example substitute the elements of the vector U_x in (3) with

$$Y_i \left\{ \left| \frac{m_0(x)}{m_0(X_i)} \right| \right\}_{\frac{1}{10}}^{10},$$

that is, clipping below $\frac{1}{10}$ and above 10, which will reduce the problem for all versions of the estimator presented in this paper. Similar "clipping" precautions were recommended for related, but less likely to appear, problems in the density estimation setting in Hjort & Glad (1995), inspired by the work of Abramson (1982) and Terrell & Scott (1992).

Remark 2. An intuitive action in order to avoid the problems mentioned above is to shift all response data Y_i a distance a, say, so that the new parametric regression function $m_0(x) + a$ does not any more intersect with the x-axis, see also Jones *et al.* (1995). The local polynomial estimators are invariant to such shifts, e.g. substituting Y_i with $Y_i + a$ produces the estimator $\tilde{m}(x,\ p) + ae_1^{\mathrm{T}}(X_x^{\mathrm{T}} Z_x X_x)^{-1} X_x^{\mathrm{T}} Z_x 1 = \tilde{m}(x,\ p) + a$, where 1 is the vector that has 1 in all n entries. For our semiparametric regression estimator, also substituting $m_0(x)$ with $m_0(x) + a$, we obtain instead

$$e_1^{\mathrm{T}}(X_x^{\mathrm{T}} Z_x X_x)^{-1} X_x^{\mathrm{T}} Z_x M + ae_1^{\mathrm{T}}(X_x^{\mathrm{T}} Z_x X_x)^{-1} X_x^{\mathrm{T}} Z_x N,$$

with the $(n \times 1)$ vectors M and N having entries

$$M_i = \left\{ Y_i \frac{m_0(x) + a}{m_0(X_i) + a} \right\} \quad \text{and} \quad N_i = \left\{ \frac{m_0(x) + a}{m_0(X_i) + a} \right\}.$$

Since $(m_0(x) + a)/(m_0(X_i) + a)$ converges to 1 as $a \to \pm\infty$, the estimator becomes more and more similar to the usual local polynomial estimator as a increases in size. Hence, a large shift of the responses solves the problem of the zeros, but at the same time reduces the effect of the parametric guide.

3. The new estimator with an estimated guide

Instead of keeping the initial choice of regression function fixed, we now allow it to belong to some parametric family of functions $m(x, \beta)$, where $\beta = (\beta_1, \beta_2, \ldots, \beta_q)^{\mathrm{T}}$ is a q-dimensional vector of parameters, to be estimated from the data by some usual estimation method. By the same arguments as in the previous section, replacing the fixed $m_0(x)$ with the fitted parametric model $m(x, \hat{\beta})$, the semiparametric regression function estimator becomes

$$\hat{m}(x,\ p) = m(x, \hat{\beta}) \cdot \hat{r}(x,\ p) = e_1^{\mathrm{T}}(X_x^{\mathrm{T}} Z_x X_x)^{-1} X_x^{\mathrm{T}} Z_x U_x, \tag{7}$$

where the vector U_x is now

$$U_x = \left(Y_1 \frac{m(x, \hat{\beta})}{m(X_1, \hat{\beta})}, \ldots, Y_n \frac{m(x, \hat{\beta})}{m(X_n, \hat{\beta})} \right)^{\mathrm{T}}. \tag{8}$$

The asymptotic properties for this generalized local polynomial regression estimator are stated below in theorem 2. Compared to the more simple situation with a fixed start, parameter estimation variability now influences the results. Anticipating the final conclusions, up to the orders of interest, there is actually no loss in precision caused by this estimation step, as we will show next.

In studying the behaviour of this estimator, we first note that the parameter estimation is possibly performed outside the parametric model conditions. For concreteness we focus on $\hat{\beta}$ being a maximum likelihood (ML) estimator. Recall that (X, Y) are generated from the smooth density $f(x)g(y|x)$. Let $f(x)g_\beta(y|x)$ denote the density under the chosen parametric assumption on the conditional expectation. Whether the parametric model is the true source of the data or not, the ML estimator $\hat{\beta}$ will aim at the value β_0 which minimizes the Kullback–Leibler distance from the true $f(x)g(y|x)$ to the suggested density $f(x)g_\beta(y|x)$,

$$d_{\mathrm{KL}} = \int\int f(x)g(y|x)\log\frac{g(y|x)}{g_\beta(y|x)}\,dy\,dx.$$

Hence β_0 is defined as the least false value of β with respect to this distance and $m(x, \beta_0)$ is the corresponding conditional mean function. If the parametric form happens to be the true one, β_0 will be the true value of the parameter. In the case of additive, normally distributed noise, the criterion for β_0 becomes especially simple. Let $g(y|x) \sim \mathcal{N}(m(x), \sigma^2(x))$ and $g_\beta(y|x) \sim \mathcal{N}(m(x, \beta), \sigma^2(x))$. Minimizing the Kullback–Leibler distance above is then equivalent to minimizing the weighted L_2-distance from the true mean function $m(x)$ to the parametric model $m(x, \beta)$,

$$d_{L_2} = \int (m(x) - m(x, \beta))^2 f(x)/\sigma^2(x)\,dx.$$

Extending to more general estimators than the ML estimator, let P denote the generating simultaneous distribution of (X, Y) and P_n the corresponding empirical distribution. Let the estimator of β be of functional form, $\hat{\beta} = T(P_n)$, where T denotes a certain functional. This $\hat{\beta}$ aims at a certain value $\beta_0 = T(P)$, the one that makes $m(x, \beta_0)$ the best approximant to $m(x)$ with respect to some distance measure d. Following the arguments in Hjort & Glad (1995), we will allow all estimators $\hat{\beta}$ that are such that $\hat{\beta} - \beta_0$ can be expressed as an average of i.i.d. zero mean variables plus remainder terms,

$$\hat{\beta} - \beta_0 = \frac{1}{n}\sum_{i=1}^{n} I(X_i, Y_i) + \frac{b}{n} + \epsilon_n, \tag{9}$$

where $I(X, Y)$ is the q-dimensional influence function

$$I(X, Y) = \lim_{\epsilon\to 0}\{T((1-\epsilon)P + \epsilon\delta_{(X,Y)}) - T(P)\}/\epsilon$$

with zero mean and finite covariance matrix. Here we write $\delta_{(X,Y)}$ for the unit point mass in (X, Y). The remaining term ϵ_n in (9) has mean $O(n^{-2})$, hence b/n is essentially the bias of the estimator.

Along with expression (9) for $\hat{\beta} - \beta_0$, we make a Taylor expansion of the elements of U_x in (8) around β_0, giving

$$\begin{aligned} Y_i\frac{m(x, \hat{\beta})}{m(X_i, \hat{\beta})} = Y_i\frac{m(x, \beta_0)}{m(X_i, \beta_0)}\{1 + (u_0(x) - u_0(X_i))^{\mathrm{T}}(\hat{\beta} - \beta_0) \\ + \tfrac{1}{2}(\hat{\beta} - \beta_0)^{\mathrm{T}}G(\hat{\beta} - \beta_0)\} + O_p(n^{-2}) \end{aligned} \tag{10}$$

where $G = (v_0(x) - v_0(X_i)) + (u_0(x) - u_0(X_i))(u_0(x) - u_0(X_i))^{\mathrm{T}}$, and $u_0(x)$ and $v_0(x)$ are

the gradient and the Hessian matrix with respect to β, respectively, of $\log m(x, \beta)$, computed in β_0. Note that G is a matrix.

Using the expansion of U_x in (10), the estimator in (7) can be rewritten as

$$\hat{m}(x,\ p) = m^*(x,\ p) + V_n(x) + \tfrac{1}{2}W_n(x) + O_p(n^{-2}), \tag{11}$$

where $m^*(x, p)$ is equal to the fixed start estimator in the previous section using $m_0(x) = m(x, \beta_0)$, and the two additional terms

$$V_n(x) = e_1^{\mathrm{T}}(X_x^{\mathrm{T}} Z_x X_x)^{-1} X_x^{\mathrm{T}} Z_x V_x$$

and

$$W_n(x) = e_1^{\mathrm{T}}(X_x^{\mathrm{T}} Z_x X_x)^{-1} X_x^{\mathrm{T}} Z_x W_x,$$

where the $(n \times 1)$ vectors V_x and W_x with entries

$$Y_i \frac{m_0(x)}{m_0(X_i)}(u_0(x) - u_0(X_i))^{\mathrm{T}}(\hat{\beta} - \beta_0), \quad i = 1, \ldots, n$$

and

$$Y_i \frac{m_0(x)}{m_0(X_i)}(\hat{\beta} - \beta_0)^{\mathrm{T}} G(\hat{\beta} - \beta_0), \quad i = 1, \ldots, n,$$

are due to the parameter estimation variability.

Theorem 2

Let $m_0(x) = m(x, \beta_0)$ be the best parametric approximation to $m(x)$, with $\beta_0 = T(P)$, and let $r(x) = m(x)/m_0(x)$. Assume $m_0 \in \mathscr{C}^{p+2}(S)$ and that $|m_0| > \delta > 0$ on R. Under condition 1, as $n \to \infty$, $h \to 0$, and $nh \to \infty$, the semiparametric estimator $\hat{m}(x, p)$ in (7) satisfies

for p odd,

$$E(\hat{m}(x,\ p)|X_1, \ldots, X_n) = m(x) + h^{p+1}\left\{\frac{m_0(x)r^{(p+1)}(x)}{(p+1)!}\right\}\left\{\int t^{p+1} K_{(p)}(t)\, dt\right\} + o_P(h^{(p+1)}),$$

for p even,

$$E(\hat{m}(x,\ p)|X_1, \ldots, X_n) = m(x) + h^{p+2}\left\{\frac{m_0(x)r^{(p+2)}(x)}{(p+2)!} + \frac{m_0(x)r^{(p+1)}(x)f'(x)}{(p+1)!f(x)}\right\} \times \left\{\int t^{p+2} K_{(p)}(t)\, dt\right\} + o_P(h^{(p+2)}),$$

and in either case

$$\mathrm{var}(\hat{m}(x,\ p)|X_1, \ldots, X_n) = (nh)^{-1}\left\{\frac{\sigma^2(x)}{f(x)}\right\}\left\{\int K_{(p)}^2(t)\, dt\right\} + o_P((nh)^{-1}).$$

Proof. Also here we present the proof for $p = 1$. In order to derive the above expressions, we

make use of the reformulation of the estimator as shown in (11) and the representation of $\hat{\beta} - \beta_0$ displayed in (9).

Hence we look to the guided estimator $\hat{m}(x, 1)$ rewritten as

$$\hat{m}(x, 1) = m^*(x, 1) + V_n(x) + \tfrac{1}{2}W_n(x) + O_p(n^{-2}),$$

and note that the first term $m^*(x, 1)$ has conditional expectation and variance as in theorem 1. The proof proceeds showing that the contributions of $V_n(x)$ and $W_n(x)$ to expectation and variance are of negligible order.

We start with $E(V_n(x)|X_1, \ldots, X_n) = e_1^{\mathrm{T}}(X_x^{\mathrm{T}} Z_x X_x)^{-1} X_x^{\mathrm{T}} Z_x S_x$ where the $(n \times 1)$ vector $S_x = E(V_x|X_1, \ldots, X_n)$ has entries

$$\frac{m_0(x)}{m_0(X_i)}(u_0(x) - u_0(X_i))^{\mathrm{T}} E(Y_i(\hat{\beta} - \beta_0)|X_1, \ldots, X_n), \quad i = 1, \ldots, n.$$

Using the representation of $(\hat{\beta} - \beta_0)$ as in (9) and the fact that $E(I(X_i, Y_i)|X_1, \ldots, X_n) = 0$ for all i, these vector elements become

$$n^{-1}s_x(X_i) = n^{-1}\frac{m_0(x)}{m_0(X_i)}(u_0(x) - u_0(X_i))^{\mathrm{T}} E(Y_i I(X_i, Y_i) + Y_i b|X_1, \ldots, X_n) + O_P(n^{-2}), \quad i = 1, \ldots, n.$$

Expanding $s_x(X_i)$, $i = 1, \ldots, n$, around x as we did for $r(X_i)$ in the proof of theorem 1 and following the same procedure, we get

$$E(V_n(x)|X_1, \ldots, X_n) = n^{-1}s_x(x) + n^{-1}h^2\left\{\frac{s_x''(x)}{2}\right\}\left\{\int t^2 K(t)\,dt\right\} + o_P(h^2/n).$$

But $s_x(x) = 0$ and hence $E(V_n(x)|X_1, \ldots, X_n) = O_P(h^2/n)$. Analogously, $E(W_n(x)|X_1, \ldots, X_n)$ can also be shown to be of negligible order $O_P(h^2/n)$, which completes the proof for the expectation.

For the variance, we start with $\mathrm{var}(V_n(x)|X_1, \ldots, X_n)$. This variance can be derived analogously to the variance of $\hat{m}(x, 1)$ in the proof of theorem 1, but now with C_x in (4) being the variance–covariance matrix of V_x with elements

$$c_x(X_i, X_j) = \frac{m_0^2(x)}{m_0(X_i)m_0(X_j)}\,\mathrm{cov}(Y_i(u_0(x) - u_0(X_i))^{\mathrm{T}} \cdot(\hat{\beta} - \beta_0), Y_j(u_0(x) - u_0(X_j))^{\mathrm{T}}(\hat{\beta} - \beta_0)|X_1, \ldots, X_n).$$

With some care, we find in this case that $c_x(X_i, X_j) = n^{-1}t_x(X_i, X_j)$ for $j \neq i$ and $c_x(X_i, X_i) = n^{-2}q_x(X_i)$, say, where both $t_x(X_i, X_j)$ and $q_x(X_i)$ involve terms of the form $(u_0(x) - u_0(X_i))^{\mathrm{T}}$. Inserting these expressions for $c_x(X_i, X_j)$ in (4), and again approximating the averages with expectations, we see that all the leading terms of $X_x^{\mathrm{T}} Z_x C_x Z_x X_x$ vanish because both $t_x(x, x)$ and $q_x(x)$ are 0. By applying the pre- and post-multiplication of $(X_x^{\mathrm{T}} Z_x X_x)^{-1}$ we prove that $\mathrm{var}\,(V_n(x)|X_1, \ldots, X_n) = O_P(h/n)$ (or even smaller).

Similar arguments can also be used to prove that $\mathrm{var}(W_n(x)|X_1, \ldots, X_n)$, and the covariances $\mathrm{cov}(m^*(x, 1), V_n(x)|X_1, \ldots, X_n)$, $\mathrm{cov}(m^*(x, 1), W_n(x)|X_1, \ldots, X_n)$, and $\mathrm{cov}(V_n(x), W_n(x)|X_1, \ldots, X_n)$ are all at least as small as $O_P(h/n)$. Better bounds can be obtained, but are not of interest here since $O_P(h/n)$ is already negligible compared to $o_P((nh)^{-1})$. The claim follows.

Here we have considered conditional properties for computational simplicity. For $p = 0$ both theorem 1 and theorem 2 are proved unconditionally in Glad (1996). For local polynomials of

degree more than 0, unconditional asymptotic properties are to our knowledge derived only for the local linear estimator, in Fan (1993). For further comments on unconditional properties of local polynomial estimators, see Seifert & Gasser (1996).

As we see by comparing theorem 2 with theorem 1, the parameter estimation has not induced, up to the order considered, any additional asymptotic bias or variance to the new estimator. Looking to the expansion in (10), this can be explained by the fact that $\hat{\beta} - \beta_0$ is small at the same time as the estimator uses only X_is which are close to x, making also $u_0(X_i)$ close to $u_0(x)$ and $v_0(X_i)$ close to $v_0(x)$.

3.1. Multivariate version

The new regression estimator can easily be extended to the multidimensional case. Avoiding new vector notations, let the predictors X_i have dimension d and let $p(x, y) = f(x)g(y|x)$ denote the joint density as before. Use for simplicity a bandwidth matrix $\mathrm{diag}\{h_1^2, \ldots, h_d^2\}$ and a product type symmetric, d-variate density K as kernel, that is,

$$K_h(X_i - x) = \prod_{j=1}^{d} \frac{1}{h_j} \mathscr{K}\left(\frac{X_{ij} - x_j}{h_j}\right).$$

Then the local linear regression estimator with a parametric guide for higher dimensions is exactly as the univariate one in (7).

The statistical properties are however changed to

$$E(\hat{m}(x, 1)|X_1, \ldots, X_n) = m(x) + \left\{\frac{\sum_{j=1}^{d} h_j^2 m_0(x) r''_{jj}(x)}{2}\right\}\left\{\int t^2 \mathscr{K}(t)\, dt\right\} + o_P\left(\sum_{j=1}^{d} h_j^2\right),$$

$$\mathrm{var}(\hat{m}(x, 1)|X_1, \ldots, X_n) = (nh_1 h_2 \cdots h_d)^{-1}\left\{\frac{\sigma^2(x)}{f(x)}\right\}\left\{\int \mathscr{K}^2(t)\, dt\right\}^d + o_P((nh_1 h_2 \cdots h_d)^{-1}).$$

The corresponding expectation of the usual local linear estimator in d dimensions is

$$E(\tilde{m}(x, 1)|X_1, \ldots, X_n) = m(x) + \left\{\frac{\sum_{j=1}^{d} h_j^2 m''_{jj}(x)}{2}\right\}\left\{\int t^2 \mathscr{K}(t)\, dt\right\} + o_P\left(\sum_{j=1}^{d} h_j^2\right),$$

while the variance is the very same as that of the new multidimensional estimator up to the order used (derived by Ruppert & Wand, 1994). Unconditional and slightly more general results for a guided multivariate Nadaraya–Watson estimator can be found in Glad (1996). Comparing the two leading bias terms, the possibility for bias reduction is essentially the same as for the univariate case, that is, we need that the parametric guiding family captures some of the features of $m(x)$, making $m_0(x)r''_{jj}(x)$ smaller in size than

$m''_{jj}(x)$. Since traditional non-parametric estimators have rather slow convergence rates in higher dimensions, we speculate that the parametric guiding idea proves even more helpful in the multivariate setting.

4. Criteria for comparison

In the comparison of approximated mean squared error (mse) and mean integrated squared error (mise) of the new and the traditional estimators, we need only to look to the leading terms of the asymptotic bias. For general p, the new estimator in (7) obtains bias reduction if the parametric model is chosen so that the inequalities (5) or (6) are fulfilled. In the following we will concentrate on the case $p = 1$. In terms of approximated mse, the parametrically guided local linear estimator is better than the local linear estimator for all x where

$$|m_0(x)r''(x)| < |m''(x)|, \tag{12}$$

a condition which applies also to the Nadaraya–Watson case if one assumes that the predictors are fixed, equally spaced or $f(x)$ is uniform. Several other kernel type estimators also have the coefficient $m''(x)$ in the leading term of asymptotic bias (Jones *et al.*, 1994), hence (12) yields a bias comparison also between the new estimator and these.

In terms of the global approximated weighted mise the new estimator is better than the local linear if its roughness functional

$$R_{\text{new}} = \int [m_0(x)r''(x)]^2 w(x)\, dx$$

is smaller than the corresponding one for the traditional estimator,

$$R_{\text{trad}} = \int [m''(x)]^2 w(x)\, dx.$$

Here $w(x) \geqslant 0$ is some weight function, typically used to downweight contributions from the boundaries. For all statements above, the extension to general odd p is straightforward. It is common to compare estimators also in terms of minimum mise values, performing a best case vs best case comparison. For the new estimator as well as for the traditional one, the smoothing parameter h^* that minimizes the approximated mise is

$$h^* = \left[\frac{\{\int K^2(t)\, dt\} \int \sigma^2(x) w(x)/f(x)\, dx}{nR\{\int t^2 K(t)\, dt\}^2} \right]^{1/5}, \tag{13}$$

where R is the corresponding roughness functional R_{new} or R_{trad} defined above. Analogous results for general p can be found in Fan & Gijbels (1996, ch. 3). The resulting minimum mise is $\text{mise}^* = \text{mise}(h^*) = c \cdot R^{1/5}$, where c is a constant. Hence the relative improvement or worsening of minimum mise can be quantified as $\text{mise}^*_{\text{new}}/\text{mise}^*_{\text{trad}} = (R_{\text{new}}/R_{\text{trad}})^{1/5}$.

In general, the bias is reduced to negligible order and the new estimator will be superior with respect to both mse and mise if the truth $m(x)$ happens to be a member of the specified parametric family $m(x, \beta)$. If $m(x)$ is in some neighbourhood around the chosen pilot class so that the parametric model captures some of the features of the form of $m(x)$, then $r(x)$ will be less rough than $m(x)$ itself and hence its second derivative will be smaller in size. In order to investigate such neighbourhoods or quantify the possible gain we need to be less general and resort to illustrating examples. In the following, two simple examples are discussed, applying a linear and an exponential parametric guide to various truths.

4.1. A linear guide

We first investigate the asymptotic bias reduction criteria (12) in the case of a linear start function $m(x, \beta)$ together with various truths. Let the true conditional mean function belong to the class of functions

$$m(x) = x \exp(ax) \tag{14}$$

for x in some interval (c, d) on R^+ and any $a \in R$. This regression function is linear when the parameter $a = 0$ and deviates more and more from the straight line as a increases or decreases. We are interested in understanding how a linear choice for the parametric start influences the bias of the new estimator for different values of a in the underlying truth. We use a linear start $m(x, \beta) = \beta x$ without intercept. The slope β is to be estimated from the data.

Assuming $g(y|x) \sim \mathcal{N}(m(x), \sigma^2(x))$ and using maximum likelihood estimation for β, the best (in the Kullback–Leibler sense) linear approximant $m_0(x) = m(x, \beta_0)$ to $m(x)$ is found by minimizing with respect to β the weighted L_2-distance, giving

$$\beta_0 = \frac{\int_c^d x^2 \exp(ax) f(x)/\sigma^2(x)\, dx}{\int_c^d x^2 f(x)/\sigma^2(x)\, dx}.$$

The correction function $r(x) = m(x)/m_0(x)$ is in this situation $r(x) = \beta_0^{-1} \exp(ax)$, giving

$$r''(x) m_0(x) = a^2 x \exp(ax).$$

Note that β_0 is no longer involved. The expression above is to be compared in absolute value with

$$m''(x) = (2a + a^2 x)\exp(ax),$$

hence the linear start induces bias reduction as long as

$$|a^2 x \exp(ax)| < |(2a + a^2 x)\exp(ax)|. \tag{15}$$

This is immediately seen to be fulfilled for all $a \geqslant 0$ for all x. Hence, the simple linear guide gives absolute improvement of the non-parametric method with respect to both (a)mse and (a)mise for any $a \geqslant 0$. For example, if we consider for x the interval (0, 1) and $a = 1$, we get $R_{\text{new}}/R_{\text{trad}} = 0.070$, implying that the minimum mise ratio is only 0.59.

For $a < 0$, we have to look separately to the cases $x < -2/a$ and $x > -2/a$. In the first case, we achieve bias reduction whenever $x < -1/a$. In the second case (15) is never true. This result is not surprising when looking more carefully to the true conditional mean function in (14) for $a < 0$. In this case the function is unimodal with its maximum in $x = -1/a$, as illustrated in Fig. 1. Hence except on an interval $(c, d) \subseteq (0, 1/a)$ it makes certainly no sense to approximate the regression function with a straight line starting in the origin, and only extremely sparse and noisy data could mislead us to such a choice.

4.2. An exponential guide

Now let the parametric start be a simple exponential function $m(x, \beta) = \beta \exp(-x)$ and let the true underlying conditional mean function be

$$m(x) = \exp(-(1 + a)x) \tag{16}$$

for x on some interval (c, d) on R^+ and some $a > -1$. This function belongs to the start

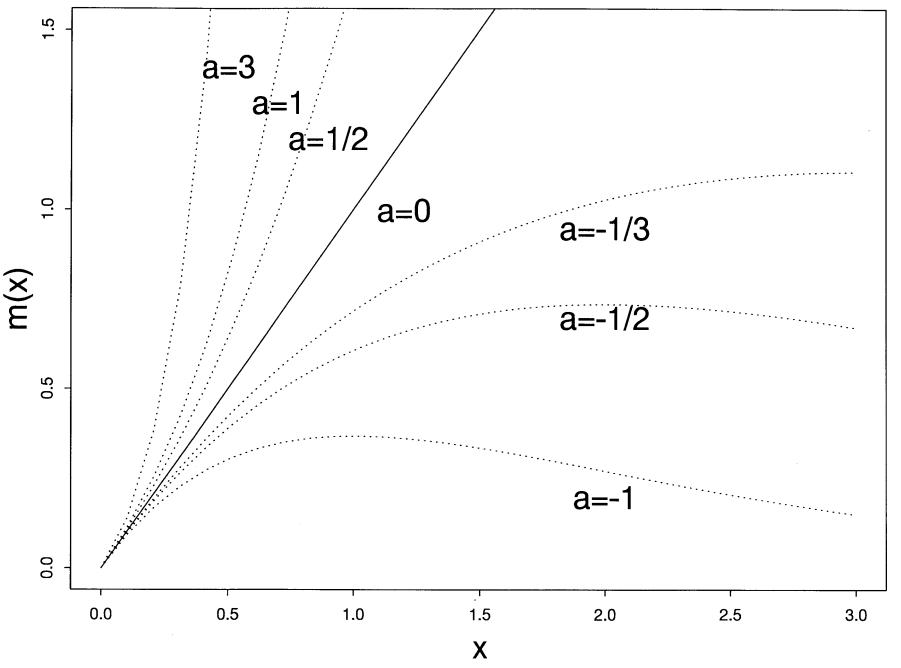

Fig. 1. The regression function $m(x)$ in (14) plotted for some different as. The solid line corresponds to the linear start. The new method with a linear start is advantageous for all regression curves with positive a. In the case of negative a the regression curve is unimodal and starts to decline at $-1/a$. The curve for $a = -1/2$, for example, has its maximum in $x = 2$. The new method with a linear start is advantageous up to this value of x.

class of simple exponentials when $a = 0$ and deviates more and more from this class as a tends to -1 and infinity, illustrated in Fig. 2.

The least squares estimator $\hat{\beta} = \sum_{i=1}^{n} Y_i \exp(-X_i)/\sum_{i=1}^{n} \exp(-2X_i)$ can be shown to aim at

$$\beta_0 = \frac{\int_c^d \exp(-(2+a)x) f(x)/\sigma^2(x)\,dx}{\int_c^d \exp(-2x) f(x)/\sigma^2(x)\,dx}$$

by using the same criteria as in the previous section. This factor disappears in the following calculations. Again we write $m_0(x) = m(x, \beta_0)$.

For calculating the asymptotic bias of the semiparametric estimator we have $r(x) = \beta_0^{-1} \exp(-ax)$, leading to

$$r''(x)m_0(x) = a^2 \exp(-(1+a)x).$$

The corresponding term for the non-parametric estimators is

$$m''(x) = (1+a)^2 \exp(-(1+a)x),$$

giving the criteria for asymptotic bias reduction of the new estimator as

$$|a^2 \exp(-(1+a)x)| < |(1+a)^2 \exp(-(1+a)x)|. \tag{17}$$

Of course, when the truth belongs to the same class as the parametric start, that is $a = 0$, the leading term of the bias of the new estimator is reduced to 0 while the strictly non-parametric goes as $\exp(-x)$.

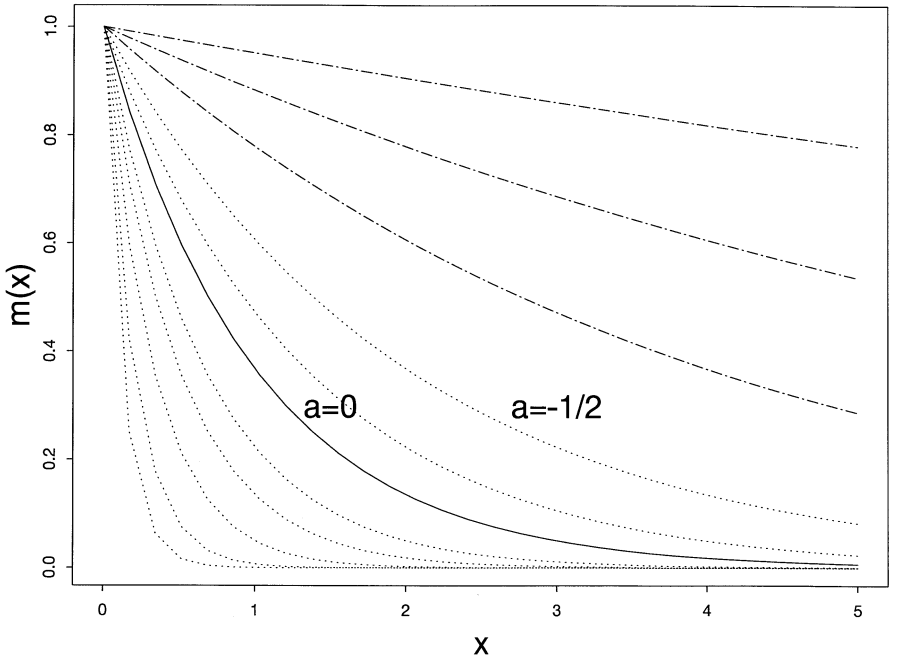

Fig. 2. The regression function $m(x)$ in (16) plotted for some different as. The solid line corresponds to the simple exponential start. Examples of regression curves that have $a > -\frac{1}{2}$ and hence obtain bias reduction are plotted with dotted lines. Curves that have $a < -\frac{1}{2}$ and hence no bias reduction are plotted with dashed-dotted lines.

The leading terms R_{new} and R_{trad} of the global mise are

$$R_{\text{new}} = a^4 \int_c^d \exp(-2(1+a)x)\, dx \tag{18}$$

$$R_{\text{trad}} = (1+a)^4 \int_c^d \exp(-2(1+a)x)\, dx. \tag{19}$$

The inequality in (17) holds for the whole range of x for any $a > -\frac{1}{2}$, and so does $R_{\text{new}} < R_{\text{trad}}$ for any interval (c, d). Letting the interval (c, d) be the whole positive line, (18) and (19) give the minimum mise ratio as $\text{mise}^*_{\text{new}}/\text{mise}^*_{\text{trad}} = (a/(1+a))^{4/5}$, which is plotted for illustration in Fig. 3 as function of a. Taking $a = 1$, for example, $R_{\text{new}}/R_{\text{trad}} = 0.063$, and the minimum mise ratio is 0.57.

For $-1 < a < -\frac{1}{2}$, the simple exponential start turns out to be too far from the truth to improve the bias. For such as $m(x)$ tends quite rapidly towards a constant line in 1, hence a plausible explanation is that these curves benefit more from a constant start, which is exactly the traditional local linear estimator. The plot of the minimum mise ratio in Fig. 3 demonstrates that there is, however, all in all very little to lose and much to gain by applying an exponential guide to data generated from a class of functions as in (16).

5. Performance on simulated data

In this section we investigate the finite sample performance of the new class of estimators based on simulations. Given a true regression function $m(x)$, we let the n design points X_i

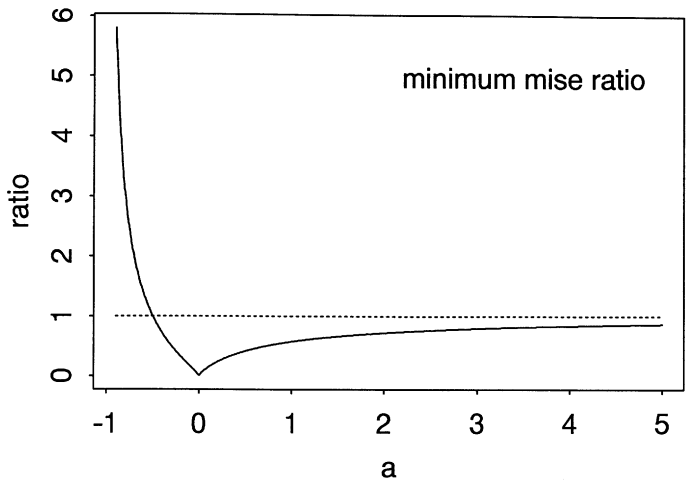

Fig. 3. The minimum mise ratio plotted as functions of a. Notice that this ratio is smaller than 1 for all $a \in [-\frac{1}{2}, \infty)$.

be drawn from a uniform density on [0, 1], and we generate the responses using additive, normally distributed noise with $\sigma^2(x) = \sigma^2$.

We consider the local linear estimator with a Gaussian kernel function and use an estimated smoothing parameter. We apply a "rule of thumb" estimate of h^* in (13) as described for example in Fan & Gijbels (1996, ch. 4.2). The weight function is taken to be $w(x) = w_0(x)f(x)$ where $w_0(x) = 1$ on [0.1, 0.9] and 0 otherwise. Hence we have used

$$\hat{h}^* = 0.776 \left[\frac{\sigma^2 \int w_0(x)\, dx}{\sum_{i=1}^{n} [\check{m}''(X_i)]^2 w_0(X_i)} \right]^{1/5},$$

where $\check{m}$ is a global fourth order polynomial fit to the data, serving as a rough approximation to $m(x)$. Note that in these experiments, we never utilize the fact that we know the design density $f(x)$.

For the new estimator, we use various forms for the parametric guide $m(x, \hat{\beta})$ and least squares parameter estimation, in combination with a local linear estimator as above. Hence the estimator has the form as in (7) with $p = 1$. We have kept here the smoothing parameter that is optimal for the usual local linear estimator, hence the comparison is slightly unfair towards the new guided estimator.

Experiments have been run with various true regression curves, various parametric guides, sample sizes ranging from 25 to 1000 and different amounts of noise. For each set of data, the different estimators are applied to obtain the corresponding estimated curves on a grid on [0, 1]. Based on 600 such realizations, we plot the average curves to visualize the biases. Furthermore, we calculate the sum over the grid of squared bias plus variance for each estimator. These mise like measures are presented along with the figures, keeping the contributions from bias and variance terms separated. We also look to the ratio of this quantity for the new estimator and the traditional one, to be less than 1.0 if the guided estimator has a better over all performance.

We present results from two types of regression curves, with a few parametric guides for each. First, the guide is guessed correctly and belongs to the true parametric family; second, the guide is obviously wrong; and finally, a more reasonable guide, though not the correct one, in assumed, based on visual inspection of the data.

The actual functions presented in this section are:

$m(x)$	$2 + \sin(2\pi x)$	$2 + x - 2x^2 + 3x^5$
σ	0.50	0.70
$m(x, \hat{\beta})$	$\hat{\beta}_0 + \hat{\beta}_1 \sin(2\pi x)$	$\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \hat{\beta}_3 x^5$
	$\hat{\beta}_0 + \hat{\beta}_1 x$	$\hat{\beta}_0 + \hat{\beta}_1 x$
	$\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \hat{\beta}_3 x^3$	$\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 \exp(x)$

The first conclusion drawn from extensive simulation experiments is that we obtain significant bias reduction even for rather rough parametric guides. For clearly misleading guides, the new estimator has a tendency to "ignore" the wrong information and behave very similarly to the purely non-parametric estimator. This corresponds to our experience in density estimation, see Hjort & Glad (1995, 1996).

The bias reduction is evident for all sample sizes, even very small ones. For extremely small n, there is a tendency to a slight increase of the variance, that, in accordance with the asymptotic results in theorem 2, vanishes when n increases.

In Fig. 4, we display results for $n = 25$ for the sine regression function. As seen, even for so few data points, the estimated sine guide improves the estimation significantly. The third degree

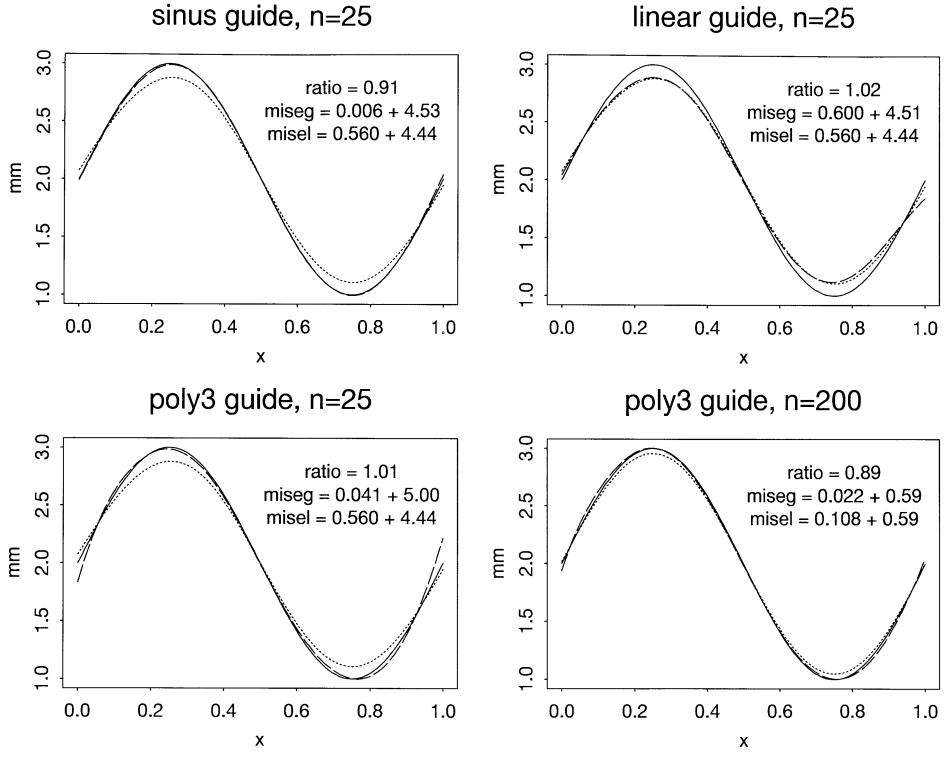

Fig. 4. Estimation of $m(x) = 2 + \sin(2\pi x)$ with different guides. The solid line is the true curve, the dotted line is the local linear estimator and the dashed line is the guided estimator. The curves are averages over 600 independent estimates. The two terms in "miseg" represent the sum of squared bias and the sum of variance, respectively, over the grid for the guided estimator. The corresponding quantities for the local linear estimator are labeled "misel", and "ratio" is miseg/misel.

polynomial guide also reduces the bias considerably, but has a slightly increased "mise" due to a somewhat larger variance. Increasing the sample size to $n = 200$, this effect disappears and the estimator behaves completely in accordance with the asymptotic theory. Note that this guide is not the correct one, but still captures many of the features of the underlying sine curve. Still in Fig. 4, we see that a linear guide has almost no effect on the estimate of the sine regression function, even if it is obviously misleading.

In Fig. 5, results of the simulations with the polynomial curve are presented for $n = 50$. Using a fifth degree polynomial as the guide, the bias becomes negligible compared to that of the local linear estimator, but due to the high level of noise in these simulations, the variance is a bit increased. With more data points, $n = 200$, this picture considerably improves. The exponential guide also has a good bias reducing effect, exhibiting only a small increase in the variance for $n = 50$, resulting in an improved "mise" value. For this regression curve, the linear guide actually leads to a slight improvement of the performance, even if the truth is far from linear.

It should be remembered that the simulation results presented here are based on rather small sample sizes, but still the guided estimator shows strong bias reducing properties. Actually, all not too unreasonable guides give significant bias reduction for all sample sizes and all levels of noise in our experiments. Furthermore, we have used a smoothing parameter that is not optimal for the new estimator. With another h, the guided approach performs even better.

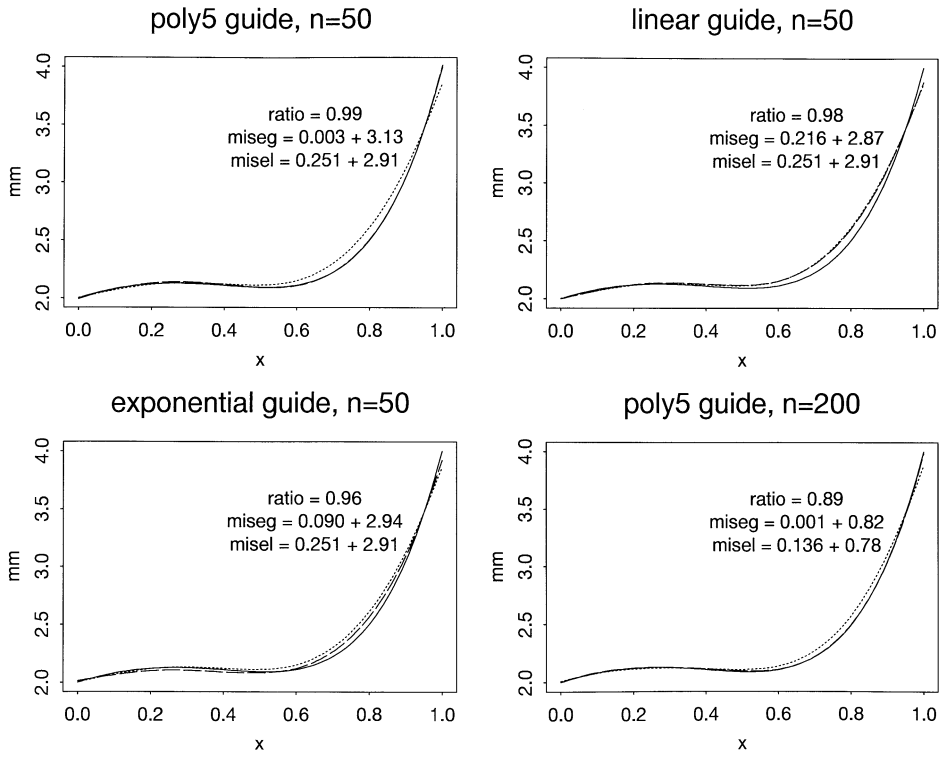

Fig. 5. Estimation of $m(x) = 2 + x - 2x^2 + 3x^5$ with different guides. The solid line is the true curve, the dotted line is the local linear estimator and the dashed line is the guided estimator. The curves are averages over 600 independent estimates. The two terms in "miseg" represent the sum of squared bias and the sum of variance, respectively, over the grid for the guided estimator. The corresponding quantities for the local linear estimator are labeled "misel", and "ratio" is miseg/misel.

Finally, we add that for real applications, more advanced parametric estimation procedures than the linear regression applied here could be more appropriate. For instance, the method of delete-knot regression splines of Smith (1982) and Breiman & Peters (1992), where knots and hence the number of parameters are reduced according to a cross-validation criteria, are likely to be useful as the parametric guide for most regression functions in practice.

Acknowledgments

I am grateful to Nils Lid Hjort for introducing me to this topic, and for his always valuable advice. I also thank Dipartimento MeMoMat at the University of Rome 'La Sapienza' for their hospitality. The author has been supported by the Research Council of Norway.

References

Abramson, I. S. (1982). On bandwidth variation in kernel estimates—a square root law. *Ann. Statist.* **10**, 1217–1223.

Breiman, L. & Peters, S. (1992). Comparing automatic smoothers (a public service enterprise). *Int. Statist. Rev.* **60**, 271–290.

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

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

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

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

Fan, J. & Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman & Hall, London.

Fan, Y. & Ullah, A. (1996). Asymptotic normality of a combined regression estimator. Manuscript, University of Windsor.

Gasser, T. & Müller, H. G. (1979). Kernel estimation of regression functions. *Smoothing techniques for curve estimation.* Lecture Notes in Mathematics **757**, 23–68. Springer Verlag, New York.

Glad, I. K. (1996). A note on unconditional properties for a parametrically guided Nadaraya–Watson estimator. Statistical Research Report, Department of Mathematics, University of Oslo, No. 19 1996. To appear in *Statist. Probab. Lett*. (1998).

Gozalo, P. & Linton, O. (1996). Using parametric information in nonparametric regression. Cowles Foundation Discussion Paper No. 1075.

Hastie, T. J. & Loader, C. (1993). Local regression: automatic kernel carpentry (with comments). *Statist. Sci.* **8**, 120–143.

Hjort, N. L. & Glad, I. K. (1995). Nonparametric density estimation with a parametric start. *Ann. Statist.* **23**, 882–904.

Hjort, N. L. & Glad, I. K. (1996). On the exact performance of a multiplicative semiparametric density estimator. Statistical Research Report, Department of Mathematics, University of Oslo, No. 8 1996. Submitted.

Jones, M. C., Davies, S. J. & Park, B. U. (1994). Versions of kernel-type regression estimators. *J. Amer. Statist. Assoc.* **89**, 825–832.

Jones, M. C., Linton, O. & Nielsen, J. P. (1995). A simple and effective bias reduction method for density and regression estimation. *Biometrika* **82**, 327–338.

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

Scott, D. W. (1992). *Multivariate density estimation: theory, practice, and visualization.* Wiley, New York.

Seifert, B. & Gasser, T. (1996). Finite-sample variance of local polynomials: Analysis and solutions. *J. Amer. Statist. Assoc.*, **91**, 267–275.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics.* Wiley, New York.

Smith, P. (1982). Curve fitting and modeling with splines using statistical variable selection techniques. NASA Contractor Report 166034.

Stone, C. J. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.* **5**, 595–620.

Terrell, G. R. & Scott, D. W. (1992). Variable kernel density estimation. *Ann. Statist.* **20**, 1236–1265.
Wand, M. P. & Jones, M. C. (1995). *Kernel smoothing.* Chapman & Hall, London.

Received November 1996, in final form November 1997

Ingrid K. Glad, Department of Mathematics, University of Oslo, P.b. 1053 Blindern, 0316 Oslo