Einbeck, Jochen; Augustin, Thomas

Working Paper

On weighted local fitting and its relation to the Horvitz-Thompson estimator

Discussion Paper, No. 465

Provided in Cooperation with:
Collaborative Research Center (SFB) 386: Statistical Analysis of discrete structures - Applications in Biometrics and Econometrics, University of Munich (LMU)

Suggested Citation: Einbeck, Jochen; Augustin, Thomas (2005) : On weighted local fitting and its relation to the Horvitz-Thompson estimator, Discussion Paper, No. 465, Ludwig-Maximilians-Universität München, Sonderforschungsbereich 386 - Statistische Analyse diskreter Strukturen, München,
https://doi.org/10.5282/ubm/epub.1834

This Version is available at:
https://hdl.handle.net/10419/31119

On weighted local fitting and its relation to the Horvitz-Thompson estimator

Jochen Einbeck*
National University of Ireland
Department of Mathematics
Galway, Ireland

Thomas Augustin†
Institut für Statistik
Ludwigstr. 33
80539 München, Germany

7th December 2005

Abstract

Weighting is a largely used concept in many fields of statistics and has frequently caused controversies on its justification and profit. In this paper, we analyze a weighted version of the well-known local polynomial regression estimators, derive their asymptotic bias and variance, and find that the conflict between the asymptotically optimal weighting scheme and the practical requirements has a surprising counterpart in sampling theory, leading us back to the discussion on Basu's (1971) elephants.

Key Words:

Bias reduction, nonparametric smoothing, local polynomial modelling, kernel smoothing, leverage values, Horvitz-Thompson theorem, stratification.

1 Introduction

What does "weighted local fitting" mean? This title seems to contain a pleonasm, since local fitting is in a certain sense always weighted, where weighting enters by means of kernel functions. More specifically, assume we are given a

*jochen.einbeck@nuigalway.ie
†augustin@stat.uni-muenchen.de

random sample $(x_1, y_1), \ldots, (x_n, y_n)$ drawn from a certain bivariate population $(X, Y) \in \mathbb{R}^2$ with mean function $m(x) = E(Y|X = x)$ and variance function $\sigma^2(x) = \mathrm{Var}(Y|X = x)$. Let $K(\cdot)$ be a kernel function and h denote the bandwidth. A local polynomial estimator (Ruppert & Wand, 1994) of degree p for m at point x is generally given by $\hat{m}(x) = \beta_0(x)$, where $\beta_0(x)$ is obtained by solving the minimization problem

$$\min_{\beta} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) \left(y_i - \sum_{j=0}^{p} \beta_j(x)(x_i - x)^j\right)^2 \tag{1}$$

w.r.t. $\beta = (\beta_0(x), \ldots, \beta_p(x))$. In particular, setting $p = 0$ leads to the Nadaraya-Watson estimator (Nadaraya, 1964), and $p = 1$ yields a local linear estimator (Fan, 1992). The kernel function $K(\cdot)$ is usually assumed to be a bounded probability density function, e.g. the Gaussian density or the Epanechnikov kernel, $K(u) = \frac{3}{4}(1 - u^2) \cdot I_{[-1,1]}(u)$. The use of a kernel function is motivated by a simple and obvious fact: Data pairs (x_i, y_i) with x_i lying near to the target value x contain more relevant information about $m(x)$ than data points being located far away from x. Note that this kind of weighting might be described as *fair* weighting: With x moving through the data, every data point (x_i, y_i) has once the chance to achieve the maximum weight $K(0)$, namely when $x = x_i$. In other words, the weighting scheme only depends on the distance between x_i and x, but not on the position of x_i itself. An *unfair* weighting scheme is obtained by introducing an additional weight function, say $\alpha(\cdot)$, in minimization problem (1), yielding

$$\min_{\beta} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) \alpha(x_i) \left(y_i - \sum_{j=0}^{p} \beta_j(x)(x_i - x)^j\right)^2 . \tag{2}$$

Several settings of $\alpha(\cdot)$ have been proposed for special situations. In the case of parametric regression, i.e. $h \longrightarrow \infty$, 'it is natural to favor observations with small variances by weighting the sum of squares' (Huet, Bouvier, Gruet & Jolivet, 1996), and the resulting weight function

$$\alpha(x_i) = 1/\sigma^2(x_i) \tag{3}$$

can be shown to be optimal in a variance-minimizing sense (see Carroll & Ruppert, 1988, for a profound treatment of this kind of weighting). For nonparametric regression, however, this does not hold, and some authors suggested to

set

$$\alpha(x_i) = f^k(x_i), \tag{4}$$

where f is the design density and k some constant. An early approach in this direction was pursued by Fan & Gijbels (1992), who additionally replaced (for $p = 1$) the fixed bandwidth h with the variable bandwidth $h/\alpha(x_i)$. The resulting weighted local estimator corresponds in the case $k = 1/4$ to a smoothing spline (Silverman, 1984) and in the case $k = 1$ to a nearest-neighbor estimator (Jennen-Steinmetz & Gasser, 1988). Fan & Gijbels (1992) showed that the asymptotically optimal weight function is proportional to $f^{1/4}(x)/\sigma^2(x)$.

In this paper, however, we concentrate on the case of a constant bandwidth h as in Einbeck, de André & Singer (2004), who proposed to set k equal to some small positive integer, e.g. $k = 1$ or 2. The aim of this choice of k was achieving robustness against outliers in the design space. Fig. 1 shows a simple example taken from the latter article. A local linear smoother (dotted line) and a weighted local linear smoother (solid; with $k = 2$) are fitted to the number of respiratory deaths of children under five as a function of SO_2 concentration, recorded in the city of São Paulo from 1994 to 1997. One observes that the unweighted curve is misleading, suggesting that the risk of respiratory death decreases for very high concentrations of SO_2. The problem of horizontal outliers (i.e. outliers in the design space) has received much less attention in the statistical literature than that of vertical outliers. One possible reason may be that the former type of outliers was frequently denied to be an outlier at all; e.g. Barnett & Lewis (1994), p. 318, argued that 'an extreme ('outlying') value in the design space of an experiment lacks the fortuitous (probabilistic) stimulus for its extremeness which we have adopted as a characteristic of outlying behavior'. This is certainly true for fixed design, but might not be the adequate point of view if the design is random, as in the example given above. We follow the usual convention in this paper and identify the term *outlier* with an *outlying response*, and write *outlying predictor* to stress that the value is outlying in the $x-$direction.

When we talk about *weighted local* fitting in this paper, the term *weighted* refers to the function $\alpha(\cdot)$ (note the semantical difference to *locally weighted* fitting,

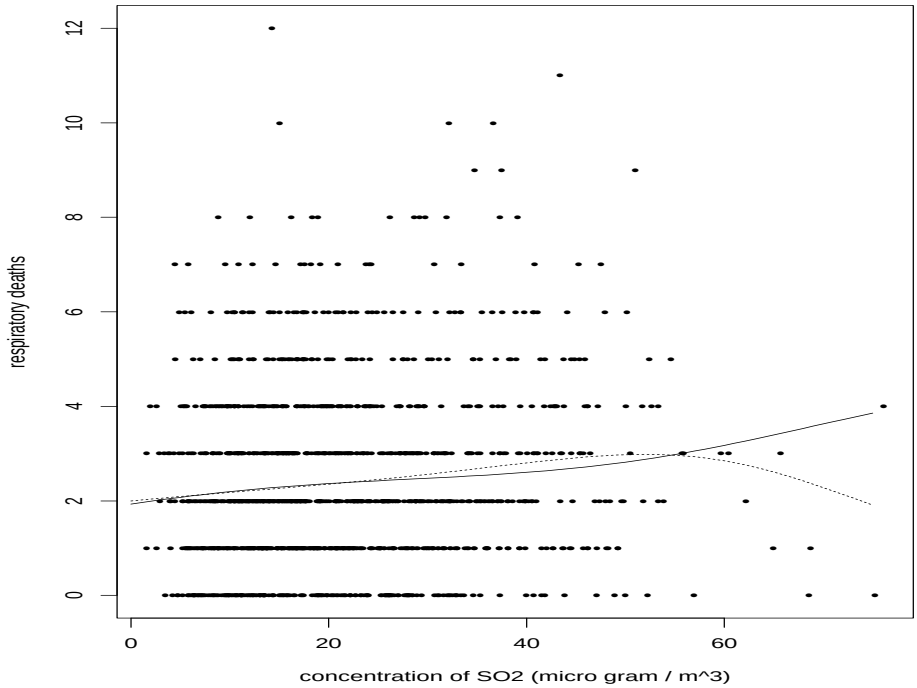

Figure 1: (Einbeck, de André and Singer) Respiratory deaths versus SO_2 concentration, local linear fit (dotted) and fit with robustness to horizontal outliers (solid).

which just refers to the use of kernel functions). The former type of weighting has indeed to be called *unfair*, since a priori some data points (X_i, Y_i) get associated to higher weights than other ones. The paper is organized as follows. In Section 2, we investigate in detail the properties of weighted local estimators obtained by minimizing (2). In particular, the asymptotic behavior is studied and an asymptotically optimal weight function is derived, which turns out to be of the form (4) with $k = -1$. In Section 3, this weight function is compared to the weights based on the setting $k = 1$, and a small simulation study is provided to give an impression of the behavior of differently weighted estimators. As similar weighting concepts are well-known from sampling theory (see e.g. Kish, 1990), we compare the findings in Section 4 with related theoretical results from this field and find surprising analogies, helping us to understand problems better. The paper finishes with the Conclusion in Section 5.

2 Properties of the weighted local smoother

In this section, we analyze the properties of the estimators

$$\hat{m}^{(j)}(x,\alpha) = j!\hat{\beta}_j(x) \tag{5}$$

for the $j-$th derivative ($0 \leq j \leq p$) of m at x, which are obtained from the minimizers $\hat{\beta}_j(x)$ of (2) according to Taylor's theorem. It is convenient to introduce matrix notation. Let therefore

$$X = \begin{pmatrix} 1 & x_1 - x & \cdots & (x_1-x)^p \\ \vdots & \vdots & & \vdots \\ 1 & x_n - x & \cdots & (x_n-x)^p \end{pmatrix}, y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$
$$W = \mathrm{diag}(K_h(x_i - x))_{1\leq i\leq n}, \quad A = \mathrm{diag}(\alpha(x_i))_{1\leq i\leq n}.$$

Then the minimization problem (2) can be written in the form

$$\min_{\beta}(y - X\beta)^T AW(y - X\beta). \tag{6}$$

The solution

$$\hat{\beta} = (X^T AWX)^{-1}X^T AWy,$$

is similar like for common local polynomial fitting (Ruppert & Wand, 1994). Then $\hat{m}^{(j)}(x,\alpha) = e_{j+1}^T\hat{\beta}$, where $e_{j+1} = (0,\ldots,0,1,0,\ldots,0)^T$, with 1 at $(j+1)^{th}$ position, serves as an estimator for $m^{(j)}(\cdot)$ at point x. For instance, for $p = 0$ one obtains the weighted local constant estimator

$$\hat{m}(x,\alpha) = \frac{\sum_{i=1}^n \alpha(x_i)K_h(x_i - x)x_i}{\sum_{i=1}^n \alpha(x_i)K_h(x_i - x)}, \tag{7}$$

where $K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$. Furthermore it is easily verified that

$$\mathrm{Bias}(\hat{\beta}|\mathbb{X}) = (X^T AWX)^{-1}X^T AWr, \tag{8}$$

where $r = (m(x_1),\ldots,m(x_n))^T - X\beta$ is the vector of the residuals of the local approximation and $\mathbb{X}$ denotes the vector of predictors $(x_1,\ldots,x_n)$. The conditional covariance matrix is given by

$$\mathrm{Var}(\hat{\beta}|\mathbb{X}) = (X^T AWX)^{-1}(X^T A^2\Sigma X)(X^T AWX)^{-1}, \tag{9}$$

where $\mathbf{\Sigma} = \mathrm{diag}(K_h^2(x_i - x)\sigma^2(x_i))$.

2.1 Asymptotical properties

We denote the kernel moments by

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

and define the matrices of kernel moments

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

With $o_P(1)$ denoting a sequence of random variables which tends to zero in probability, we have the following proposition:

Proposition 1. *Under assumptions (i) to (v) (see Appendix A) one gets for* $h \longrightarrow 0$

$$\text{Bias}(\hat{\beta}|\mathbb{X}) = h^{p+1} H^{-1} \left[\beta_{p+1} S^{-1} c_p + h b^*_\alpha(x) + o_n\right] \tag{10}$$

and

$$\text{Var}(\hat{\beta}|\mathbb{X}) = \frac{\sigma^2(x)}{f(x)nh} H^{-1} \left[S^{-1}S^*S^{-1} + h V^*_\alpha(x) + o_n\right] H^{-1} \tag{11}$$

where $H = diag(1, h, \ldots, h^p)$, $o_n = o_P(h) + O_P\left(\frac{1}{\sqrt{nh}}\right)$,

$$b^*_\alpha(x) = \left(\frac{\alpha'(x)}{\alpha(x)} + \frac{f'(x)}{f(x)}\right) \beta_{p+1} \left(S^{-1}\tilde{c}_p - S^{-1}\tilde{S}S^{-1}c_p\right) + \beta_{p+2} S^{-1}\tilde{c}_p \tag{12}$$

and

$$\begin{aligned} V^*_\alpha(x) &= \left(2\frac{\sigma'(x)}{\sigma(x)} + 2\frac{\alpha'(x)}{\alpha(x)} + \frac{f'(x)}{f(x)}\right) S^{-1}\tilde{S}^*S^{-1} - \\ &- \left(\frac{\alpha'(x)}{\alpha(x)} + \frac{f'(x)}{f(x)}\right) \cdot \left(S^{-1}\tilde{S}S^{-1}S^*S^{-1} + S^{-1}S^*S^{-1}\tilde{S}S^{-1}\right). \end{aligned} \tag{13}$$

A sketch of the proof is provided in the appendix. The formulas given in this proposition reduce to the expressions provided in Fan, Gijbels, Hu & Huang (1996) in the special case $\alpha(\cdot) \equiv 1$. Note that the leading bias and variance terms are independent of $\alpha(\cdot)$. This can also be seen in the following proposition, which is obtained from Proposition 1 using formula (5):

Proposition 2. *Let* $h \longrightarrow 0$ *and* $nh \longrightarrow \infty$. *Under assumptions (i) to (v)*

$$\text{Var}(\hat{m}^{(j)}(x,\alpha)|\mathbb{X}) = e_{j+1}^T S^{-1}S^*S^{-1}e_{j+1} \frac{j!\sigma^2(x)}{f(x)nh^{1+2j}} + o_p\left(\frac{1}{nh^{1+2j}}\right) \tag{14}$$

and

$$\begin{aligned}&\text{Bias}(\hat{m}^{(j)}(x,\alpha)|\mathbb{X}) = \\ &\quad = \quad e_{j+1}^T S^{-1} c_p \frac{j!}{(p+1)!} m^{(p+1)}(x) h^{p+1-j} + o_P(h^{p+1-j}) \end{aligned} \tag{15}$$

hold.

Both formulas are the same as those for local polynomial fitting (Fan & Gijbels, 1996, Theorem 3.1). Note that application of Propositions 1 and 2 needs some care when symmetric kernels are used, as in this case the odd kernel moments and hence some kernel moment matrix products vanish. In particular, for the variance formulas, the expressions $e_{j+1}^T S^{-1} \tilde{S} S^{-1} S^* S^{-1} e_{j+1}$, $e_{j+1}^T S^{-1} S^* S^{-1} \tilde{S} S^{-1} e_{j+1}$ and $e_{j+1}^T S^{-1} \tilde{S}^* S^{-1} e_{j+1}$ are trivially zero for any choice of p and j, while the expression $e_{j+1}^T S^{-1} S^* S^{-1} e_{j+1}$ is never trivially zero.

The situation is more complicated for the bias expression, where $e_{j+1}^T S^{-1} c_p$ is zero for $p-j$ even, while $e_{j+1}^T S^{-1} \tilde{c}_p$ and $e_{j+1}^T S^{-1} \tilde{S} S^{-1} c_p$ are zero for odd values of $p-j$. This special behavior motivates to formulate the bias for symmetric kernels in a separate proposition, taking the deeper expansion of the bias (12) into account:

Proposition 3. *Let $h \longrightarrow 0$ and $nh^3 \longrightarrow \infty$. Under assumptions (i) to (vi) we get for $p-j$ odd*

$$\begin{aligned}&\text{Bias}(\hat{m}^{(j)}(x,\alpha)|\mathbb{X}) = \\ &\quad = \quad e_{j+1}^T S^{-1} c_p \frac{j!}{(p+1)!} m^{(p+1)}(x) h^{p+1-j} + o_P(h^{p+2-j}) \end{aligned} \tag{16}$$

and for $p-j$ even

$$\begin{aligned}&\text{Bias}(\hat{m}^{(j)}(x,\alpha)|\mathbb{X}) = \\ &\quad = \quad e_{j+1}^T \frac{j!}{(p+1)!} \left[\left(\frac{\alpha'(x)}{\alpha(x)} + \frac{f'(x)}{f(x)} \right) \left(S^{-1} \tilde{c}_p - S^{-1} \tilde{S} S^{-1} c_p \right) m^{(p+1)}(x) + \right. \\ &\qquad \left. + S^{-1} \tilde{c}_p \frac{m^{(p+2)}(x)}{p+2} \right] h^{p+2-j} + o_P(h^{p+2-j}). \end{aligned} \tag{17}$$

The second formula provided in Proposition 3 is remarkable, because it shows that in this special case the leading term is *not* independent of $\alpha(\cdot)$. This gives the chance to reduce the bias. Note that the augend in the squared bracket in (17) vanishes for

$$\frac{\alpha'(x)}{\alpha(x)} + \frac{f'(x)}{f(x)} = 0,$$

and this differential equation is solved for

$$\alpha_{opt}(x) = c\frac{1}{f(x)}, \tag{18}$$

with $c \in \mathbb{R}\backslash\{0\}$. This result is in various aspects surprising: Fan (1992) and Fan & Gijbels (1996) argued that the order p of the polynomial should be chosen such that $p - j$ is odd, since in this case the estimators are design-adaptive, meaning that the asymptotic bias does not depend on the design density and its derivatives. Estimators based on even values of $p - j$ are not design-adaptive and should consequently be avoided. Regarding (17) and (18), we see that the disturbing term depending on the density can be completely eliminated, if only $f(\cdot)$ is known and the weighting $\alpha(\cdot) = \frac{1}{f(\cdot)}$ is applied. Thus, the role of the function $\alpha(\cdot)$ is in fact to manipulate the influence of the design density. In practice, certainly, $f(\cdot)$ is mostly unknown, but may be substituted by a suitable density estimate $\hat{f}(\cdot)$.

2.2 Leverage values

The second remarkable point about the asymptotically optimal weights (18), which suggest to set $k = -1$ in (4), is that this seems to be in contrast to the proposal $k = 1$ from Einbeck, de André & Singer (2004) mentioned in the introduction. Does there exist some foundation for the latter setting as well? There is at least a heuristic one. Recall that the hat matrix L of a smoother $\hat{m}$ is defined by

$$\begin{pmatrix} \hat{m}(x_1) \\ \vdots \\ \hat{m}(x_n) \end{pmatrix} = Ly$$

The influence or leverage values l_i are the diagonal elements of L and can be interpreted as a measure of influence of a design point on the estimated function evaluated at this design point (for details about influence values see Huber (1981) and Hampel, Ronchetti, Rousseeuw & Stahel (1986)). Let us consider for simplicity the manipulated Nadaraya-Watson estimator (7). The leverage values of this estimator are given by

$$l_i = \frac{K_h(0)\alpha(x_i)}{\sum_{j=1}^{n} K_h(x_j - x)\alpha(x_j)} = \frac{K(0)}{h}\frac{\alpha(x_i)}{\hat{f}_\alpha(x_i)}, \tag{19}$$

where

$$\hat{f}_\alpha(x) = \frac{1}{n}\sum_{j=1}^{n} K_h(x_j - x)\alpha(x_j)$$

may be seen as a weighted kernel density estimate at point x. As illustrated by Loader (1999) in Fig. 2.6, the influence values of a local fit rise strongly near the boundary, which frequently falls together with regions having sparse design. From (19) we see that the leverage values are constant iff

$$\alpha(\cdot) = \hat{f}_\alpha(\cdot). \tag{20}$$

Though this formula is recursive and the weight function $\alpha(\cdot)$ appears again in the density estimate, it unveils that the weight function $\alpha(\cdot)$ plays a stabilizing role for the leverage values if it is chosen proportional to the design density. This gives some motivation for the setting $k = 1$ in (4). We illustrate this in Fig. 2 by means of a simulated data set of size $n = 50$ with beta(0.5,2)-distributed design and normally distributed errors ($\sigma = 0.3$) added to the function $y = \sqrt{x}$. As can be see from the plot in the top, the leverage values for an unweighted Nadaraya-Watson estimator rise strongly near the right boundary. Setting the weights proportional to the inverse estimated density, i.e. $k = -1$, this effect is even stronger, whereas the leverages are nearly constant for $k = 1$. For a local polynomial fit of second order these differences are not as pronounced, but the tendency is still observable, as can be seen from Fig. 2 (bottom).

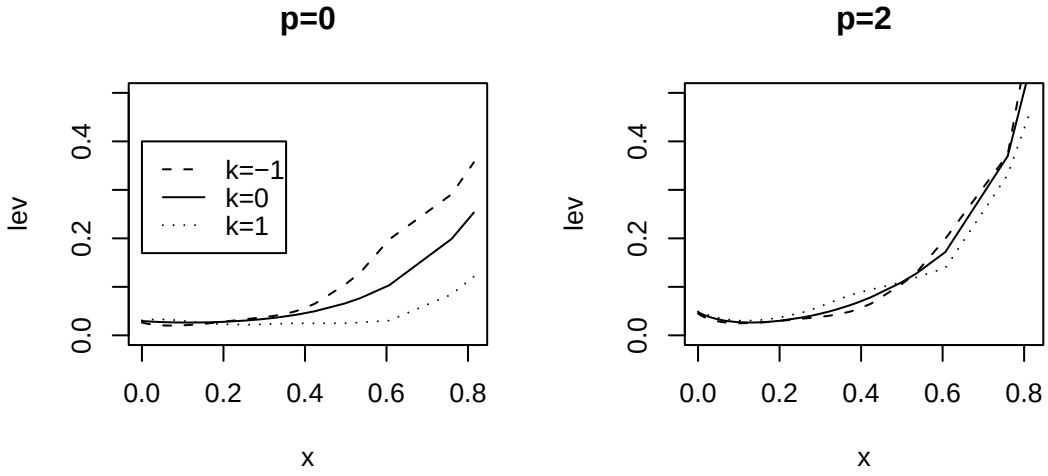

Figure 2: Leverage values for beta-distributed data ($n = 50$) for different polynomial degrees and weighting schemes.

It should be noted that extreme design points with high leverages have earlier attracted some attention in the theory of parametric regression; see e.g. the classical work by Hampel, Ronchetti, Rousseeuw & Stahel (1986), pp. 307 ff, for an overview on this research. An important parametric regression estimator based on downweighting those points is the Mallows-estimator (Mallows, 1979).

2.3 Behavior at the boundary

We have at this point two weighting schemes, which are both in some (different!) sense optimal, or at least plausible. Where is the contradiction, or is there any contradiction at all? In order to answer this question, we firstly observe from Fig. 2 that the high leverage points motivating the setting $k = 1$ also share another property: They are situated near the right boundary. However, the asymptotic results presented above concern interior points, i.e. fixed points in the interior of $f(\cdot)$. When x is a boundary point, the asymptotic behavior is different. Let us therefore take a more thorough look at the asymptotic properties of boundary points. We assume without loss of generality that the density f has a bounded support $[0, 1]$ and that f is right continuous at 0 for a left boundary point and left continuous at 1 for a right boundary point. 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 Proposition 1 and 2; the only difference is that the kernel moments μ_j and ν_j have to be replaced by

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

in case of a left boundary point, and analogously in case of a right boundary point. These kernel moments never vanish, irrespectively of whether the kernel is symmetric or not. We formulate the result in Proposition 4 for the case of a left boundary point, and omit details of the proof.

Proposition 4. *For $h \longrightarrow 0$ and $nh \longrightarrow \infty$ one gets at a left boundary point $x = ch$*

$$\mathrm{Var}(\hat{m}^{(j)}(x, \alpha)|\mathbb{X}) = e_{j+1}^T S_c^{-1} S_c^* S_c^{-1} e_{j+1} \frac{j!\sigma^2(0+)}{f(0+)nh^{1+2j}} + o_P\left(\frac{1}{nh^{1+2j}}\right),$$

and

$$\begin{aligned}&\mathrm{Bias}(\hat{m}^{(j)}(x,\alpha)|\mathbb{X}) = \\ &\quad = \quad e_{j+1}^T S_c^{-1} c_{p,c} \frac{j!}{(p+1)!} m^{(p+1)}(0+) h^{p+1-j} \; + o_P(h^{p+1-j}),\end{aligned} \tag{21}$$

where $c_{p,c} = (\mu_{p+1,c}, \ldots, \mu_{2p+1,c})^T$ *and* $S_c = (\mu_{j+l,c})_{0 \le j,l \le p}$.

In this situation, the kernel moment matrix $e_{j+1}^T S_c^{-1} c_{p,c}$ is never trivially zero. Thus, the first order approximation of the bias does not depend on $\alpha(\cdot)$, and hence the considerations leading to (18) are no longer valid for a boundary point, implying that the results (18) and (20) cannot be offhandly compared. Practically, this observation is not yet very useful, as every data set consists of interior and boundary points and needs a weight function that serves them all. In the following section, we try to work out guidelines when either setting is recommendable.

3 Discussion of different weighting schemes

When looking for a practical weight selection rule, there is one apparent and tempting idea which one might have in this connection. The weighting scheme $\alpha(\cdot) \sim f(\cdot)$ was originally introduced to robustify against outlying predictors, which is, as one might argue, rather a finite sample problem, suggesting the simple rule: Use $\alpha(\cdot) \sim f(\cdot)$ for small sample sizes, and the asymptotically optimal weights $\alpha(\cdot) \sim 1/f(\cdot)$ for large sample sizes.

3.1 A tutorial on the influence of outlying predictors

To investigate this, we consider in a tutorial manner a sample with underlying function $y = \sqrt{x}$ and beta-distributed design generated as in Section 2.2. In Fig. 3 we provide exemplarily two simulated data sets with $n = 50$ (left), and two further data sets with $n = 1000$ (right) simulated data points. The two data sets in the top are situations where either no relevant outlying predictors are present, or, if they are, their associated responses are distributed roughly symmetrically around the underlying function. In this case, the asymptotically optimal weights give indeed an excellent fit, nearly indistinguishable from the fit with constant

weights. Weighting with the estimated density at some target point x gives too much weight to the *previous* observations compared to the next ones, so that the estimate oversteers. The matter is different in the situations in the bottom, where outlying high leverage points are present. Here the asymptotic weighting scheme can produce a heavy bias in sparse data regions, whereas the robustified version stays comparatively near to the underlying function. Hence, there is no guarantee at all that either weight function improves the fit, as outlying leverage points may or may not occur for any sample size and (bounded or unbounded) design.

It is, of course, a question of definition if extreme design points as generated in the right column of Fig. 3 have still to be called *outlying* predictors. Unfortunately, *"there is no generally accepted definition of what constitutes an outlier."* (Gather & Becker, 1997). Traditionally, outliers are seen as data points generated from some kind of 'contaminating' distribution, which differs from the target distribution (see e.g. Barnett & Lewis, 1994). A more modern viewpoint, brought up by Davies & Gather (1993), is to consider data points as outliers if they are far enough away from the center of the distribution of the data cloud, regardless from which distribution they are generated. For instance, for any sequence $0 < \gamma_n < 1$ the γ_n *outlier region* of the $N(\mu, \sigma^2)$ distribution is defined by

$$\mathrm{out}(\gamma_n, \mu, \sigma^2) = \{x : |x - \mu| > z_{1-\gamma_n}\sigma, \}$$

where $\gamma_n = 1 - (1-\gamma)^{1/n}$ is selected such that the probability that *no* observation falls in the outlying region is equal to $1 - \gamma$. According to this definition, the number of outlying predictors can even increase with the sample size. This seems to be counterintuitive, but is in conformity with the observations drawn from Fig. 3, where we observed no (horizontal) outliers in the sense of Davies and Gather for $n = 50$, but two outlying predictors in the bottom right picture for $n = 1000$. The beginning of the γ_n-outlying region of the beta(0.5,2) distribution for $\gamma = 0.2$, with $n = 50$ and $n = 1000$, respectively, is symbolized by a vertical line in Fig. 3. The concept of outlier regions is similar in spirit to the 'hard robustification' rule suggested by Einbeck, de André & Singer (2004).

The tutorial showed us that yet no clear statement can be made. A simulation

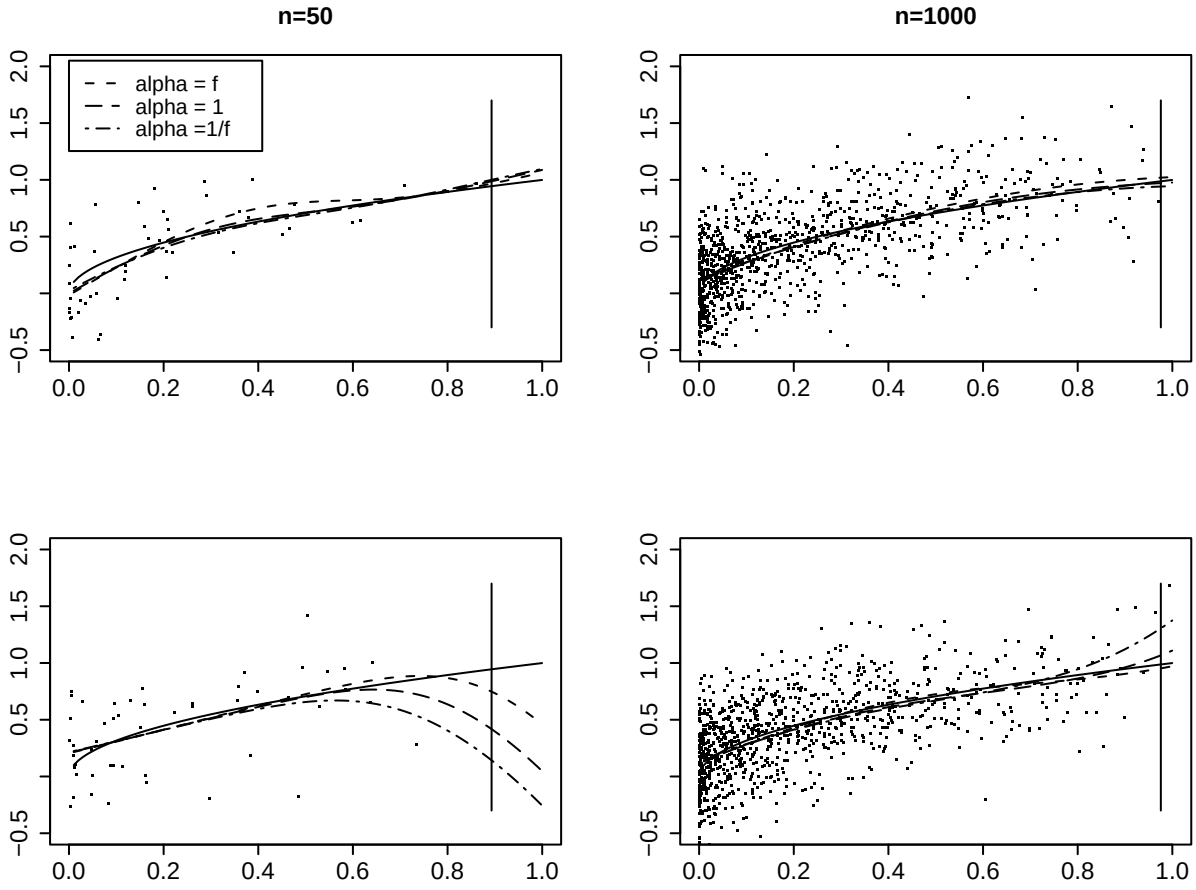

Figure 3: Selected examples for the behavior of local estimators with $p = 0$ for sample size $n = 50$ (left) and $n = 1000$ (right) with weights $\alpha = \hat{f}$, $\alpha \equiv 1$ and $\alpha = 1/\hat{f}$. The predictors follow a beta(0.5,2) distribution. The true function $y = \sqrt{x}$ is indicated by a solid line. Vertical lines indicate the beginning of the γ_n- outlying region ($\gamma = 0.2$) at 0.893 and 0.976, respectively.

study is evidently called for, and we give the results in the following.

3.2 Simulation study

For data sets of size $n = 50$ and $n = 1000$, each 1000 replicates were generated as above. The choice of the error criterion needs some care in this case. Taking the average squared error as e.g. in Hart & Yi (1998),

$$ASE = \frac{1}{n}\sum_{i=1}^{n}(\hat{m}(x_i) - m(x_i))^2,$$

might overrepresent regions with dense design. Alternative choices are the integrated squared error (ISE) as used in Fan (1992) or its robust version, the integrated absolute error (IAE, Gentle, 2002, p. 146), defined by

$$\int_0^1 \ell(\hat{m}(x) - m(x))\, dx,$$

with loss function $\ell(z) = z^2$ and $\ell(z) = |z|$, respectively, where integration is performed numerically over the whole density domain, hence giving equal weight to high density and sparse regions. A variety of other criteria exist; see Fahrmeir & Tutz (2003), p. 190, for an overview. We will work representatively with the three choices outlined above, ensuring that the found results are not a particular feature of a certain error criterion. The results of the simulation study are shown in Fig. 4 for $n = 50$ (left column) and $n = 1000$ (right column) and the criteria IAE, ISA, and ASE (from top to bottom), with weighting schemes $\alpha(\cdot) = f(\cdot)$, $\alpha(\cdot) \equiv 1$ and $\alpha(\cdot) = 1/f(\cdot)$ (i.e. $k = 1, 0, -1$; from left to right within the boxplots). We distinguish two cases: In the left two columns, the density was estimated applying the kernel density estimator

$$\hat{f}(\cdot) = \frac{1}{ng}\sum_{i=1}^{n} K\left(\frac{x_i - x}{g}\right), \tag{22}$$

where the bandwidth g was selected for each simulated data set anew using Silverman's (1986, p. 48) bandwidth selector, as also proposed in Einbeck, de André & Singer (2004). In the right columns, the true (known) density was used in the weight functions $\alpha(\cdot)$.

The result in the first column is as expected. For all error criteria, there seems to be some evidence that the robust weights are superior. The second column is

alarming: For a higher sample size, the asymptotic result is even getting worse, and the robust weights stay superior. This confirms our concerns uttered in Section 3.1 that the problem of outlying predictors does not disappear with increasing sample size, but rather gains in power. This is even more remarkable as we did not assume at all in this study that the outlying predictors are in some sense ill-behaving compared to the rest of the data – all data points are simulated from the same model, and the y-values associated with the outlying predictors are not necessarily outlying in y-direction. We note at this occasion that the data set used in the introduction is actually of length $n = 1067$, giving one more example that the usefulness of this kind of weighting is not restricted to small sample sizes.

When the true density is used, however, the asymptotic weights perform – for either sample size – much better, though they never succeed to be the 'winning' weight. We return to this important observation in the next section after our look at sampling theory.

We have to stress at this point that the general picture might be different in other situations. We did a large number of simulations with different underlying functions, sample sizes, error variances, and design densities, and observed that sometimes the winning weights tended to be more on the robust, and less frequently more on the asymptotic side. It is not within the scope of this paper to give a general statement about this, therefore the simulation study is provided here just with one exemplary function. The general picture, however, was in most simulations similarly disillusioning as above as far as our initial hypothesis is concerned: Though there is a - rather small - tendency that the asymptotic weights perform better with increasing sample size, they still might give a terrific result for large sample sizes, as their success depends dramatically on the accuracy of the density estimate, and on the existence of outlying predictors with high leverages.

We seem to be not very far from where we started. Hoping to understand things better, we next take a deeper look at sampling theory, where similar theoretical results and similar practical problems, and the confusions arising from them, have already been discussed for a long time, without having much impact on

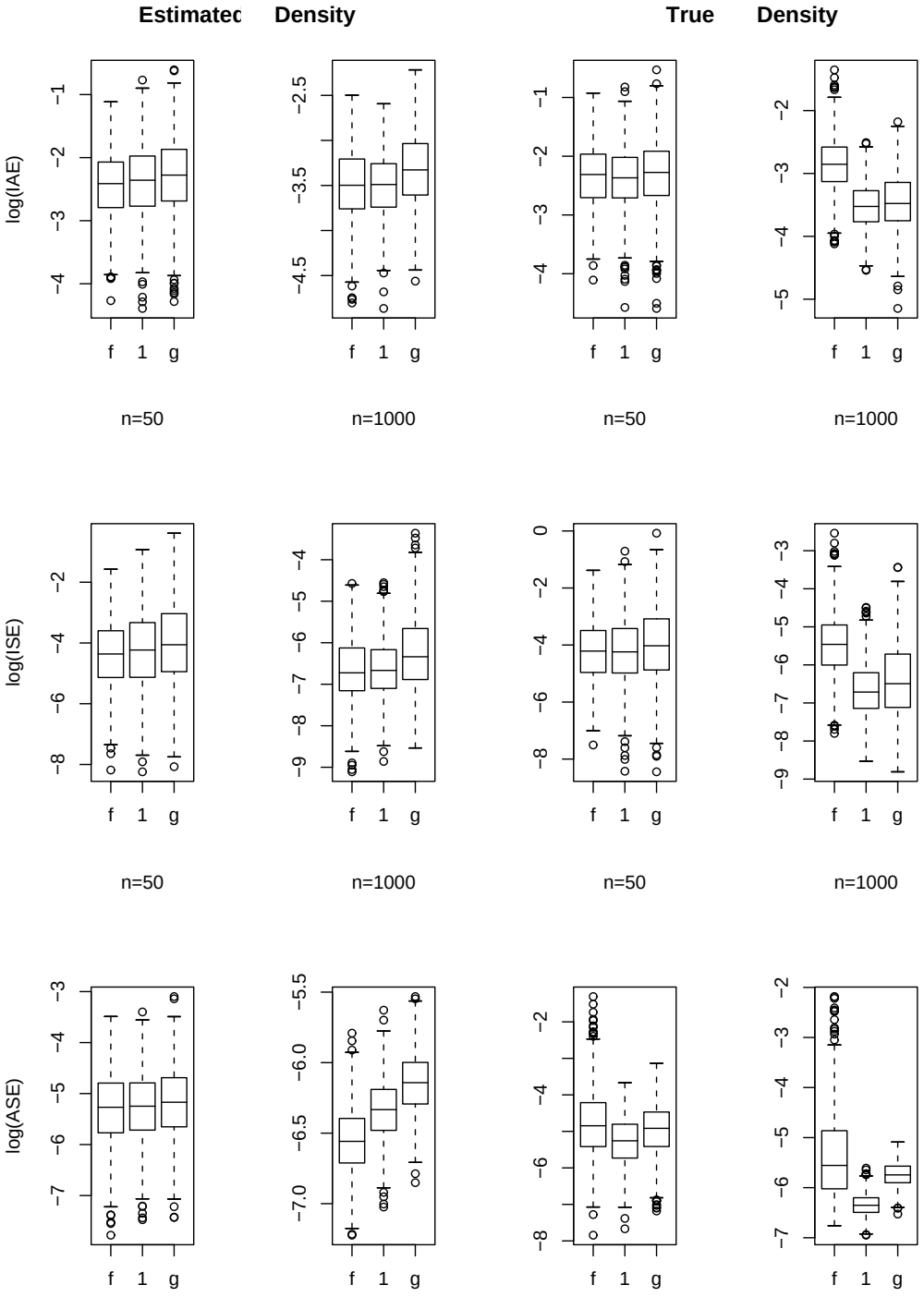

Figure 4: Weighted local regression with $p = 0$: Boxplots of $\log(IAE)$ (top), $\log(ISE)$ (middle), and $\log(ASE)$ (bottom) over 1000 simulated data sets, each with weight functions $\alpha = f, \alpha = 1$, and $\alpha = g \equiv 1/f$, for estimated (left columns) and true (right columns) densities. Note that the boxplots have differing scales, as not the absolute values are of interest, but rather the differences between weighting schemes.

other areas of statistics.

4 Relation to sampling theory

Weighting is a widely used concept in sampling theory. There exist a large variety of reasons and methods for weighting a sample, see Kish (1990) and Gabler, Hoffmeyer-Zlotnik & Krebs (1994) for overviews.

4.1 From stratification to weighted local smoothing

One of the most important reasons for weighting is stratification, where the population is divided *a priori*, i.e. before the sample is taken, into several groups, called strata, which are assumed to be more or less 'homogeneous within and heterogeneous between'. The main reasons for stratification are variance reduction or to 'produce larger samples for separate domains, usually for smaller domains' (Kish, 1990). If the proportions assigned to the strata do not meet the proportions in the population, keeping the bias small requires to weigh the strata accordingly.

We give a simple example to illustrate this. Assume one is interested in the average income of the supporters of a specific soccer team, and that it is known from some source that the target population $Y_1, \ldots, Y_N$ consists of proportions $P_m = 0.95$ men and $P_w = 0.05$ women. A sample of size $n = 500$ shall be collected in the stadium at a certain matchday. As one fears that there may be very few female spectators in a simple random sample, leading to a high variance of the estimator of this subpopulation mean, and, hence, of the target population mean (see e.g. Brewer, 2002, p. 34), one stratifies the population in a male and a female stratum, with fixed sample proportions $p_m = 0.7$ and $p_w = 0.3$, respectively. From the corresponding simple random samples within the strata, say $y_1, \ldots, y_{np_m}$ and $y_{np_m+1}, \ldots, y_n$, one calculates the means $\bar{y}_m = \sum_{i=1}^{np_m} y_i$ and $\bar{y}_w = \sum_{i=np_m+1}^{n} y_i$ for the two strata separately, assigns weights $\alpha_m = P_m/p_m = 95/70$ and $\alpha_w = P_w/p_w = 5/30$ to the observations obtained

from men resp. women, and finally computes

$$\bar{y} = \alpha_m \cdot p_m \cdot \bar{y}_m + \alpha_w \cdot p_w \cdot \bar{y}_w = \frac{1}{n} \left(\sum_{i=1}^{np_m} \alpha_m y_i + \sum_{i=np_m+1}^{n} \alpha_w y_i \right). \tag{23}$$

Obviously, using the terminology from Section 1, this kind of weighting is *unfair*, since it assigns to observations stemming from men generally a higher weight (95/70) than for those coming from women (5/30).

For a more profound analysis of this example, we introduce the factor

$$x_i = \begin{cases} 1 & \text{if observation } y_i \text{ is taken from a man,} \\ 2 & \text{if observation } y_i \text{ is taken from a woman.} \end{cases}$$

Then the individual weights of the observations may be written as $\alpha(x_1), \ldots,$ $\alpha(x_n)$, with $\alpha(x_i) = \alpha_m$ for $i \leq np_m$, and $\alpha(x_i) = \alpha_w$ otherwise. A local estimator $\bar{y}(x)$ may be interpreted as the estimator for the mean in stratum x ($x \in \{1, 2\}$). Defining the discrete kernel

$$K(x_i, x) = \begin{cases} 1 & x_i = x \\ 0 & \text{otherwise,} \end{cases} \tag{24}$$

we have

$$\bar{y}(x) = \frac{\sum_{i=1}^{n} \alpha(x_i) K(x_i, x) y_i}{\sum_{i=1}^{n} \alpha(x_i) K(x_i, x)}. \tag{25}$$

This corresponds in character to the weighted Nadaraya-Watson-estimator (7) (note that in this simple case the weights cancel out, as they are constant within each stratum, and only one stratum has non-zero kernel weights). Thus, the estimators within the strata correspond formally to the local estimators from Section 2 in the case $p = 0, j = 0$. Certainly, the asymptotic results from Section 2 do not apply directly here, as the x_i in the given example are discrete and fixed. However, let us consider a situation with a high number H of strata, denoted by ordered real numbers $k_\ell \in \mathbb{R}, \ell = 1, \ldots, H$, $k_\ell < k_m$ ($\ell < m$), with stratum proportions $P_1, \ldots, P_H$. Let the random variable X describe the event that an arbitrarily chosen sample observation stems from a certain stratum, i.e.

$$P(X = x) = \begin{cases} P_1 & \text{for} \quad x = k_1 \\ \vdots & \qquad \vdots \\ P_H & \text{for} \quad x = k_H \end{cases}$$

Applying a similar idea as in Gasser & Müller (1984), we set $s_\ell = (k_\ell + k_{\ell+1})/2$ $(1 \leq \ell \leq H-1)$, $s_0 < k_1$, and $s_H > k_H$. One gets the histogram

$$f_H(x) = \sum_{\ell=1}^{H} P_\ell \cdot 1_{\{s_{\ell-1} \leq x < s_\ell\}} / \sum_{\ell=1}^{H} P_\ell (s_\ell - s_{\ell-1})$$

for the distribution of X. If now the number of strata H tends to infinity, then the variable X loses its meaning as a discrete stratum indicator and simply represents the real axis. The series of histograms $f_H(x)$ converges then to a probability density function $f(x)$, which can be interpreted as selection probability distribution for the independent variable. Provided that the assumptions in Appendix A hold, in particular that $\alpha(\cdot)$ is smooth, and using an appropriate continuous kernel instead of (24), we can now apply the asymptotics from Section 2 on (25). Thus, the asymptotically optimal weighting scheme (18), suggesting to weight with the inverse density function, should have some relevance in the sampling context as well.

4.2 Weighted local smoothing and the Horvitz-Thompson estimator

It turns out that this is indeed the case, and that there exists already a well-known theoretical result in this direction. From a population $Y_1, \ldots, Y_N$ we draw without replacement a sample of length n. Suppose the population total $Y = \sum_{i=1}^{N} Y_i$ is to be estimated. We define the random variable δ_i, indicating whether unit i has been sampled, by

$$\delta_i = \begin{cases} 1 & \text{unit } i \text{ in sample} \\ 0 & \text{unit } i \text{ not in sample} \end{cases}$$

Horvitz & Thompson (1952) showed that among all linear estimators of the form

$$\hat{Y} = \sum_{i=1}^{N} \alpha_i \delta_i Y_i \tag{26}$$

the Horvitz-Thompson (HT) estimator

$$\hat{Y}_{HT} = \sum_{i=1}^{N} \delta_i \frac{Y_i}{\pi_i}, \tag{27}$$

is the only unbiased estimator for Y, where π_i is the probability that the i-th element is drawn in any of the n draws. Thus, in other words, the estimation is best w.r.t. the bias when the observations are weighted with the inverse selection probability. In the special case of stratification, the selection probability for an element stemming from the ℓ-th stratum is given by

$$\pi_\ell = \frac{np_\ell}{NP_\ell} = \frac{n_\ell}{N_\ell}, \tag{28}$$

where n_ℓ and N_ℓ are the size of $\ell-$th stratum in the sample and in the population, respectively (see e.g. Kish, 1965, p. 92). This is just the intuitive weighting applied in (23). DuMouchel & Duncan (1983) linked this concept to parametric regression by applying weights inversely proportional to (28) in a minimization problem of type (2) in the special case $h \longrightarrow \infty$.

For the interpretation of these results, recall that (18) means that the bias is minimized when the observations are weighted with the inverse density, while Horvitz and Thompson showed that the bias is minimized when weighting with the inverse selection probability. As the density of the independent variable in a regression problem may be considered as its selection probability distribution (and is even identical in case of a designed experiment!), this is essentially the same message. Hence, one might consider (18) as an asymptotic and nonparametric version of Horvitz-Thompson's theorem.

We illustrate this point more clearly in the following table:

Estimator	Bias minimized for	Interpretation
Horvitz-Thompson	$\alpha_i = 1/\pi_i$	π_i = selection probability of unit i,
in particular, stratification	$\alpha_\ell = 1/\pi_\ell \sim P_\ell/p_\ell$	Adaption from stratum to population proportions
weighted local, p even	$\alpha(x_i) \sim 1/f(x_i)$	$f(x_i)$ = design density at point x_i

Another important remark has to be made in this connection: Often, one notices only after the survey that the data consists of several groups. In this case, one can resort to post-stratification, where one stratifies the sample *a posteriori* in

several groups and then handles it as if it was selected a priori from different strata. Given that one knows the true strata proportions in the population, then weighting can be applied straightforwardly, and is widely used in practice, though its methodological legitimation is much fewer acknowledged (Alt & Bein, 1994). The problem is that in this case the values p_ℓ and hence $\alpha_\ell = P_\ell/p_\ell$ are not fixed, but random, and HT's theorem does not hold for random weights. Saying it sharply as Diekmann (2003), p. 366, for post-stratified samples 'it cannot be statistically justified at all that the weighted sample is less biased' than then the unweighted one. Nevertheless, it is frequently successfully applied – see Brewer (2002), p. 29ff, for an example.

This brings us back to the problem discussed in the previous section. When replacing the true design density $f(\cdot)$ with an estimated one, $\hat{f}(\cdot)$, the asymptotic results do not apply either, and the asymptotic weights (18) are not any more optimal. In this sense, using the estimated density as weight function for local smoothing is the counterpart to applying HT-weights on a post-stratified sample. Thus, it is not surprising that the simulation gave better results when the true density was applied. In contrast, the motivation given for the leverage-stabilizing weights in (19) was explicitly based on the *estimated* density. Hence, it is not surprising either that in this case the estimated density led to better results than the true density, as observed, at least for the presented example, in Section 3.2.

4.3 Once more, Basu's elephants

Hence, the theoretical results for weighted local smoothing and weighted sampling indeed meet each other and have the same interpretation. As a consequence, it is not surprising that a similar discussion as in Section 3 can be given for the HT estimator. Indeed, in the last decades there has been some confusion concerning the general applicability of the HT estimator. This confusion was provoked by Basu (1971) in his famous elephant fable: A circus owner plans to ship 50 adult elephants and therefore needs a rough estimate of their total weight. As weighing elephants is not so easy, the owner intuitively plans to weigh only one elephant and to multiply the result with 50. To decide which

elephant should be weighed, he consults the circus statistician, who assigns a selection probability of 99/100 to a previously determined elephant ('Samba') which from a previous census is known to have about the average weight of the herd. All other elephants obtain the weight 1/4900, including the elephant 'Jumbo' who is biggest of all. If Samba was now selected, its weight would have to be multiplied with 100/99 according to Horvitz-Thompson, and if Jumbo was selected, his large weight would even have to be multiplied with 4900 to get the 'best linear unbiased estimator' of the total weight. Certainly, after having given this advice, the circus statistician is sacked.

Considerations of this type led some authors to formulate statements as 'Basu's counter-example destroys frequentist sample survey theory' (Lindley, 1996). Where is actually the problem with Basu's fable? Horvitz & Thompson (1952) state that if

$$\pi_i = nY_i/Y, \tag{29}$$

the estimator $\hat{Y}$ has zero variance and the sampling will be optimal. Obviously, the probabilities in the fable are far from optimality in the sense of (29). Kish (1990) notes that 'increased variances can result from weighting ... when the selection probabilities are not optimal at all', and also Rao (1999) warns that the HT estimator 'can lead to absurd results if the π_i are unrelated to the Y_i'. Though HT's theorem can reduce the bias of an estimate *given* the inclusion probabilities, it may produce useless estimates if they are unfortunately chosen. Nevertheless, Rao judged Lindley's statement as being 'far from the truth', since HT's estimator proves to be most useful e.g. in the context of ratio estimation, when a second variable X_i is used to construct selection probabilities which are correlated to the Y_i. In Basu's example, a way out for the unfortunate circus statistician would have been to take e.g. the known elephant weights X_i from the previous census, and to set $\pi_i = nX_i/X$, where X was the total weight of the herd measured at that time (Koop, 1971, Brewer, 2002, p.63).

Though the confusions about Basu's fable have been solved at the latest with Rao's (1999) article and its subsequent discussion, it is still interesting to take a look at the rejoinder of Basu's (1971) essay, in which he vehemently denied that the 'unrealistic sampling plan' was responsible for the failure of the

Horvitz-Thompson estimator. Basu defended, in contrary, the circus statistician's sampling plan, as it ensures a *representative* sample, which would not have been guaranteed using Koop's average of ratios estimator. Instead, he gives the responsibility for the useless result entirely to the Horvitz-Thompson estimator itself, 'being a method that contradicts itself by alloting weights to the selected units that are inversely proportional to their selection probabilities. The smaller the selection probability of a unit, that is, *the greater the desire to avoid selecting the unit*, the larger the weight that it carries when selected.' Basu did not conform himself to the fact that one has to choose the probabilities adequately, and in some sense, he is right. What does one do, for instance, if no auxiliary variable X_i is available to construct a ratio estimator, or if one gets a sample, selected with 'wrong' selection probabilities, and now one has to work with it? Basu touches here exactly the problem that we have in the smoothing context. There, the π_i correspond to the $f(x_i)$, which are in the most cases inherent to the observed data or subjectively determined by the experimenter, but are not designed to meet a certain optimality criterion (Applying the bias-minimizing weights (18), one easily verifies that the variance term (13) vanishes if

$$f(x) = c_1 \cdot \sigma(x),$$

($c_1 \in \mathbb{R} \setminus \{0\}$), which is then the analogous formula to (29) and leads to weights not far from (3). However, we do not want to overvalue this result, as $V_\alpha^*(x)$ is just a second-order term). One can formulate Basu's dilemma somewhat more general: Statistical theory suggests to choose weights inversely proportional to the selection probability (distribution). This however makes the estimator extremely sensitive to 'undesired' or extreme observations (which correspond to the *outlying predictors* in the terminology of Section 3 and to 'Jumbo' in Basu's fable), if their selection probability is small. This will be almost always the case in the smoothing context, and will occur likely in the sampling context if one assigns the probabilities with the goal of representativity in mind.

We provide an other example for this dilemma, showing that, even when the weighting scheme seems to be obvious, one should not use it thoughtlessly. Survey samples of the German population are usually based on the 'ADM-Design': In the first step a region is chosen, in the second one a household via

random route, and in the third one a person in a household (e.g. Wendt, 1994). Since only one person is chosen in every household, persons living in large households have a smaller probability to be selected in the sample as persons in small households. According to HT the observations have to be weighted with their inverse selection probability, which implies that an observation stemming from an 8-persons-household has to be weighted with the factor 8. However, coming back to the discussion in Section 3: As 8-person-households rarely exist in our society; they would, in some sense, correspond to 'outlying predictors'. In addition, it can be expected that people living in such households are likely to show in some aspects different behavior than the rest of the population (this corresponds to the $y-$values associated with the outlying predictor). Can we rely on this information in a way that we give it eight times the weight of an observation obtained from a person living alone? This is the same problem as in Section 3, where we observed partly terrific results when applying the asymptotically optimal weight on data sets with outliers. Weighting in these situations has to be performed at least with care, and the influence of outlying observations on the estimates have to be checked. Similar warnings have been given in the context of design-based sampling by Alt & Bein (1994) and in Brewer (2002), p. 32 ff.

For the sake of completeness it should be noted that the situation is similar with stratification: Stratification is introduced to reduce the variance, and weighting with the inverse selection probability is then performed to reduce the bias – but might in turn lead to an increase of variance. It is well known (e.g. Kish, 1965) that that the variance is minimized when the strata are designed in a way that $\pi_\ell \equiv n_\ell/N_\ell$ is proportional to S_ℓ, where S_ℓ is the standard deviation per element in the $\ell - th$ stratum.

5 Conclusion

We have so far studied the properties of weighted local smoothers and derived an asymptotically optimal and a heuristic weighting scheme. By means of a simulation study and by resorting to sampling theory, we tried to get some practical guidelines for the choice of a weight function. The intuitively straight-

forward idea to rely on the sample size turned out to be rather misleading. It seems to play some role if the design density is known or estimated. However, it should be noted that even when employing the true design density, the asymptotic weights could not compete with the simple constant weights, though the results were in this case already a good part better than for the estimated density. Furthermore, it will be beyond common sense to suggest to base the choice of the weight function not on the available data itself, but rather on the degree of accuracy which one has for the distribution of the design points.

From our look at sampling theory we have learned that there seems to be a general dilemma with weighting procedures. If one applies the theoretical bias-minimizing weights, the estimates may get highly sensitive to outlying predictors, extreme design points, undesired observations, or howsoever the statistician in his particular field might want to call them.

As a conclusion, we have to admit that looking for an objective criterion for automatic weight selection seems to be the wrong way to approach the problem. However, a more subjective viewpoint is helpful. The asymptotical result (18) confirms the statement by Hastie & Loader (1993), who called an endpoint *'the most informative observation'* when fitting at this endpoint. Einbeck, de André & Singer (2004) added that this holds only when this point can be considered as *'as reliable as in the interior'*. This is a crucial point. Any kind of robust estimation implies that one is *not willing to trust* a certain group of data points (in this case the outlying predictors and its associated $y-$values), whereas the asymptotic result is – as HT – certainly based on full reliance on the information content of *all* data points, including outlying predictors. Hampel, Ronchetti, Rousseeuw & Stahel (1986), p. 308, already go in a similar direction when considering, in the parametric setting, *'extreme design points (which might be wrong)'*. It should however be noted that the notion of unreliability that we have in mind is somewhat more general: Beyond the extreme design points themselves, the responses associated with them might be unreliable (regardless of being outlying or not); and even if both design points and responses have to be assumed to be correct, unreliability may simply stem from the fact that there are very few observations available in an outlying region of the design space, as it is the case in the example in Fig. 1.

To formulate it again and clearly: If there is some reason to distrust some group of outlying predictors, the robust weights (4), with $k = 1$, are a reasonable choice and do their job. Otherwise, one should better stay with the usual constant weights (i.e. $k = 0$), as the asymptotically optimal weights behave disproportionately hazardous, and therefore cannot be generally recommended for practical use. For asymmetric kernels or odd values of $p - j$, e.g. a local linear estimator with $p = 1$ and $j = 0$, the effect of $\alpha(\cdot)$ vanishes asymptotically anyway.

We finally would like to encourage to look for Basu's elephants beyond the scope of smoothing and sampling – there exist a variety of other statistical concepts where weighting is performed (e.g. missing data, boosting, neural networks), and it is to expect that similar theoretical results and the related practical pitfalls appear in those areas as well.

A Assumptions

(i) The kernel K is a continuous density function having compact support;

(ii) $f(x) > 0$, $f(\cdot)$ is continuously differentiable in a neighborhood of x;

(iii) $\alpha(x) \neq 0$, $\alpha(\cdot)$ is continuously differentiable in a neighborhood of x;

(iv) $\sigma^2(x) > 0$, $\sigma^2(\cdot)$ is continuously differentiable in a neighborhood of x;

(v) $m(\cdot)$ is $p + 2$ times continuously differentiable in a neighborhood of x;

(vi) The kernel K is symmetric.

B Proof of Proposition 1

The proof is kept shortly since it follows mainly the lines of the corresponding proof for local polynomial modeling, see Fan, Gijbels, Hu & Huang (1996). Let

$w_i = K_h(x_i - x)$ and

$$r_{n,j} = \sum_{i=1}^{n} \alpha(x_i) w_i (x_i - x)^j; \qquad R_n = (r_{n,j+l})_{0 \le j,l \le p};$$
$$r^*_{n,j} = \sum_{i=1}^{n} \alpha^2(x_i) \sigma^2(x_i) w_i (x_i - x)^j; \quad R^*_n = (r_{n,j+l})_{0 \le j,l \le p}.$$

Then $R_n = X^T A W X$ and $R^*_n = X^T A^2 \Sigma X$.

Bias:

Using standard asymptotics reveals that

$$r_{n,j} = nh^j (f_\alpha(x)\mu_j + h f'_\alpha(x)\mu_{j+1} + o_n), \tag{30}$$

where $f_\alpha(x) = \alpha(x) f(x)$ and $o_n = o_P(h) + O_P\left(\frac{1}{\sqrt{nh}}\right)$, and thus

$$R_n = nH[f_\alpha(x) S + h f'_\alpha(x) \tilde{S} + o_n] H \tag{31}$$

holds. Then, using Taylor's expansion and equation (8), we get

$$\text{Bias}(\hat{\beta}|\mathbb{X}) = R_n^{-1} \left[\beta_{p+1} d_n + \beta_{p+2} \tilde{d}_n + o_P(\tilde{d}_n) \right], \tag{32}$$

where $d_n = (r_{n,p+1}, \ldots, r_{n,2p+1})^T$ and $\tilde{d}_n = (r_{n,p+2}, \ldots, r_{n,2p+2})^T$. We use the fact that $(B + hC)^{-1} = B^{-1} - hB^{-1}CB^{-1} + O(h^2)$ to calculate

$$R_n^{-1} = \frac{1}{n} H^{-1} \left[\frac{1}{f_\alpha(x)} S^{-1} - h \frac{f'_\alpha(x)}{f^2_\alpha(x)} S^{-1} \tilde{S} S^{-1} + o_n \right] H^{-1}. \tag{33}$$

Plugging (33) into (32), and substituting (30) into the vectors d_n and $\tilde{d}_n$, yields (10) via some simple matrix algebra, taking into account that

$$\frac{f'_\alpha(x)}{f_\alpha(x)} = \frac{\alpha'(x)}{\alpha(x)} + \frac{f'(x)}{f(x)}.$$

Variance:

Similar like (31) we find that

$$R^*_n = \frac{n}{h} H[s_\alpha(x) S^* + h s'_\alpha(x) \tilde{S}^* + o_n] H, \tag{34}$$

where $s_\alpha(x) = \sigma^2(x)\alpha^2(x) f(x)$. By substituting (34) and (33) in

$$\text{Var}(\hat{\beta}|\mathbb{X}) = R_n^{-1} R^*_n R_n^{-1}$$

we derive (11) by applying matrix algebra.

References

Alt, C. and Bein, W. (1994). Gewichtung, ein sinnvolles Verfahren in der Sozialwissenschaft? In S. Gabler, J. H. P. Hoffmeyer-Zlotnik, & D. Krebs (Eds.), *Gewichtung in der Umfragepraxis*, pp. 124–140. Opladen, Germany: Westdeutscher Verlag.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data (3rd Ed.)*. John Wiley.

Basu, D. (1971). An essay on the logical foundations of survey sampling, Part 1 (with discussion). In V. P. Godambe & D. A. Sprott (Eds.), *Foundations of Statistical Inference*, pp. 203–242. Toronto: Holt, Reinhart and Winston.

Brewer, K. (2002). *Combined Survey Sampling Inference*. London: Arnold.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. London: Chapman and Hall.

Davies, L. and Gather, U. (1993). The identification of multiple outliers. *J. Amer. Statist. Assoc.* **88**, 782–792.

Diekmann, A. (2003). *Empirische Sozialforschung, 10th edition*. Hamburg: Rowohlts Enzyklopädie.

DuMouchel, W. and Duncan, G. J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *J. Amer. Statist. Assoc.* **78**, 535–543.

Einbeck, J., de André, C. D. S., and Singer, J. M. (2004). Local smoothing with robustness against outlying predictors. *Environmetrics* **15**, 541–554.

Fahrmeir, L. and Tutz, G. (2003). *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer Verlag.

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

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

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

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

113–127.

Gabler, S., Hoffmeyer-Zlotnik, J. H. P., and Krebs, D. E. (1994). *Gewichtung in der Umfragepraxis*. Opladen, Germany: Westdeutscher Verlag.

Gasser, T. and Müller, H.-G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.* **11**, 171–185.

Gather, U. and Becker, C. (1997). Outlier identification and robust methods. In G. S. Maddala & C. R. Rao (Eds.), *Handbook of Statistics*, pp. 123–141. Amsterdam: Elsevier Science.

Gentle, J. E. (2002). *Elements of Computational Statistics*. New York: Springer.

Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (1986). *Robust statistics: The approach based on influence functions*. New York: Wiley.

Hart, J. D. and Yi, S. (1998). One-sided cross-validation. *J. Amer. Statist. Assoc.* **93**, 620–631.

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

Horvitz, D. G. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47**, 663–685.

Huber, P. (1981). *Robust Statistics*. New York: Wiley.

Huet, S., Bouvier, A., Gruet, M.-A., and Jolivet, E. (1996). *Statistical Tools for Nonlinear Regression*. New York: Springer.

Jennen-Steinmetz, C. and Gasser, T. (1988). A unifying approach to nonparametric regression estimation. *J. Amer. Statist. Assoc.* **83**, 1084–1089.

Kish, L. (1965). *Survey Sampling*. New York: Wiley.

Kish, L. (1990). Weighting: Why, when and how? In *ASAProc. of the Section on Survey Research Methods*, Alexandria, VA, pp. 121–130. Amer. Statist. Assoc.

Koop, J. C. (1971). Comment on: D. Basu, An essay on the logical foundations of survey sampling, Part 1. In V. P. Godambe & D. A. Sprott (Eds.), *Foundations of Statistical Inference*, pp. 236–238. Toronto: Holt, Reinhart and Winston.

Lindley, D. V. (1996). Letter to the editor. *Amer. Statist.* **50**, 197.

Loader, C. R. (1999). *Local Regression and Likelihood*. New York: Springer.

Mallows, C. L. (1979). Robust methods – some examples of their use. *Ann. Statist.* **33**, 179–184.

Nadaraya, E. A. (1964). On estimating regression. *Theory Prob. Appl.* **9**, 141–142.

Rao, J. N. K. (1999). Some current trends in sample survey theory and methods. *Sankhyã* **61**, 1–57.

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

Silverman, B. W. (1984). Spline smoothing: the equivalent variable kernel method. *Ann. Statist.* **12**, 898–916.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman & Hall.

Wendt, B. (1994). Das ADM-Stichproben - System. In S. Gabler, J. H. P. Hoffmeyer-Zlotnik, & D. Krebs (Eds.), *Gewichtung in der Umfragepraxis*, pp. 124–140. Opladen, Germany: Westdeutscher Verlag.