ON DESIGN-WEIGHTED LOCAL FITTING AND ITS RELATION TO THE HORVITZ-THOMPSON ESTIMATOR

Jochen Einbeck[1] and Thomas Augustin[2]

[1]*Durham University and* [2]*University of Munich (LMU)*

Abstract: Weighting is a widely used concept in many fields of statistics and has frequently caused controversies on its justification and benefit. In this paper, we analyze design-weighted versions of the well-known local polynomial regression estimators, derive their asymptotic bias and variance, and observe that the asymptotically optimal weights are in conflict with (practically motivated) weighting schemes previously proposed in the literature. We investigate this conflict using theory and simulation, and find that the problem has a surprising counterpart in sampling theory, leading us back to the discussion on the Horvitz-Thompson estimator and Basu's (1971) elephants. In this light one might consider our results as an asymptotic and nonparametric version of the Horvitz-Thompson theorem. The crucial point is that bias-minimizing weights can make estimators extremely vulnerable to outliers in the design space and have therefore to be used with particular care.

Key words and phrases: Bias reduction, Horvitz-Thompson estimator, kernel smoothing, leverage values, local polynomial modelling, nonparametric smoothing, stratification.

1 Introduction

This paper studies design-weighted local fitting procedures and identifies a counterpart in sampling theory. First, let us note that with "design-weighted local fitting" we do not mean locally weighted fitting in the sense of ordinary kernel weighting. To make this difference clear, assume we are given a random sample $(x_1, y_1), \ldots, (x_n, y_n)$ drawn from a bivariate population $(X, Y) \in \mathbb{R}^2$ with mean function $m(x) = E(Y|X = x)$ and variance $\sigma^2(x) = \mathrm{Var}(Y|X = x)$. Let $K(\cdot)$ be a kernel function and $h > 0$ denote the bandwidth. A local polynomial estimator (Ruppert & Wand, 1994) of degree p for m at point x is generally given by

$\hat{m}(x) = \hat{\beta}_0(x)$, where $\hat{\beta}_0(x)$ 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 the target value x contain more relevant information about $m(x)$ than data points located far away from x. 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

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

where some data points (x_i, y_i) are associated a priori with higher weights than other ones. When we refer to *design-weighted local* fitting in this paper, the term *design-weighted* refers to the function $\alpha(\cdot)$.

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 density of the design variable X 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).

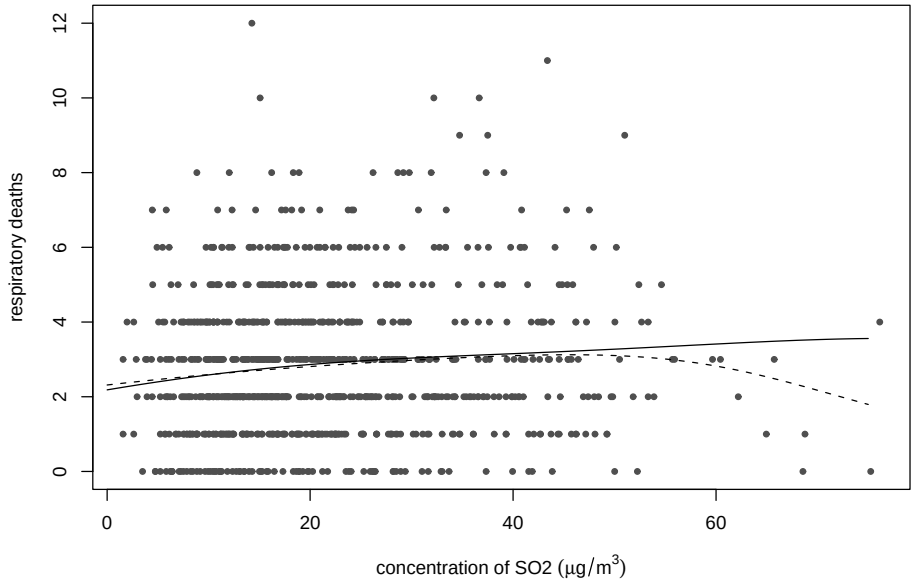

Figure 1: (Einbeck, André and Singer) Respiratory deaths versus SO_2 concentration, local linear fit (dotted) and fit with robustness to horizontal outliers (solid).

Here we concentrate on the case of a constant bandwidth h as in Einbeck, André & Singer (2004), who proposed to set k equal to some small positive integer, e.g. $k = 1$ or 2. This choice of k aims to achieve 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 design-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 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 'x_i are observational, unlike "designed" situations with fixed x_i' (Rousseeuw & van Zomeren, 1990), as in the example given above.

Accepting this, it is still debatable what characterizes an outlying design point, a question which is certainly of relevance for vertical outliers, too. 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 'contaminating' distribution, which differs from the target distribution (see e.g. Barnett & Lewis, 1994). A different 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 of the distribution from which 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 $\text{out}(\gamma_n, \mu, \sigma^2) = \{x : |x - \mu| > z_{1-\gamma_n}\sigma\}$, where $\gamma_n = 1 - (1-\gamma)^{1/n}$ is selected such that the probability that *no* observation falls into the outlying region is equal to $1 - \gamma$. It is this notion of outlyingness that we adopt in this paper.

The paper is organized as follows. In Section 2, we investigate in detail the properties of design-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 illustrate 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

We analyze the properties of the estimators

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

for the $j-$th derivative $(0 \le j \le p)$ 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

$$X = \begin{pmatrix} 1 & x_1 - x & \cdots & (x_1-x)^p \\ \vdots & \vdots & & \vdots \\ 1 & x_n - x & \cdots & (x_n-x)^p \end{pmatrix}, \; y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

$W = \mathrm{diag}(K_h(x_i - x))_{1\le i\le n}$, $A = \mathrm{diag}(\alpha(x_i))_{1\le i\le n}$, with $K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$. Then the minimization problem (2) can be written in the form $\min_\beta (y-X\beta)^T AW(y-X\beta)$, and the solution $\hat{\beta} = (X^TAWX)^{-1}X^TAWy$ is similar like for common local polynomial fitting (Ruppert & Wand, 1994). Hence, $\hat{m}^{(j)}(x,\alpha) = e_{j+1}^T\hat{\beta}$, 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

$$\hat{m}(x,\alpha) = \frac{\sum_{i=1}^n \alpha(x_i)K_h(x_i-x)y_i}{\sum_{i=1}^n \alpha(x_i)K_h(x_i-x)}. \tag{6}$$

Furthermore it is easily verified that

$$\mathrm{Bias}(\hat{\beta}|\mathbb{X}) = (X^TAWX)^{-1}X^TAWr, \tag{7}$$

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 $(x_1,\ldots,x_n)$. The conditional covariance matrix is given by

$$\mathrm{Var}(\hat{\beta}|\mathbb{X}) = (X^TAWX)^{-1}(X^TA^2\Sigma X)(X^TAWX)^{-1}, \tag{8}$$

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

2.1 Asymptotical properties

We denote the kernel moments by $\mu_j = \int_{-\infty}^{\infty} u^j K(u)\,du$ and $\nu_j = \int_{-\infty}^{\infty} u^j K^2(u)\,du$ and define the matrices of kernel moments

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

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

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

$$\begin{aligned} \mathrm{Bias}(\hat{\beta}|\mathbb{X}) &= h^{p+1} H^{-1} \left[\beta_{p+1} S^{-1} c_p + h b^*_\alpha(x) + o_n\right], &(9)\\ \mathrm{Var}(\hat{\beta}|\mathbb{X}) &= \frac{\sigma^2(x)}{f(x)nh} H^{-1} \left[S^{-1}S^*S^{-1} + h V^*_\alpha(x) + o_n\right] H^{-1} &(10) \end{aligned}$$

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

$$\begin{aligned} b^*_\alpha(x) &= \left(\frac{\alpha'(x)}{\alpha(x)} + \frac{f'(x)}{f(x)}\right)\beta_{p+1}\left(S^{-1}\tilde{c}_p - S^{-1}\tilde{S}S^{-1}c_p\right) + \beta_{p+2}S^{-1}\tilde{c}_p\,, &(11)\\ V^*_\alpha(x) &= \left(2\frac{\sigma'(x)}{\sigma(x)} + 2\frac{\alpha'(x)}{\alpha(x)} + \frac{f'(x)}{f(x)}\right) S^{-1}\tilde{S}^*S^{-1} &(12)\\ &- \left(\frac{\alpha'(x)}{\alpha(x)} + \frac{f'(x)}{f(x)}\right)\cdot\left(S^{-1}\tilde{S}S^{-1}S^*S^{-1} + S^{-1}S^*S^{-1}\tilde{S}S^{-1}\right). \end{aligned}$$

A sketch of the proof is provided in the appendix. The formulas given in this proposition reduce to the expressions provided by 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) one gets*

$$\begin{aligned} \mathrm{Var}(\hat{m}^{(j)}(x,\alpha)|\mathbb{X}) &= e_{j+1}^T S^{-1}S^*S^{-1}e_{j+1}\frac{(j!)^2\sigma^2(x)}{f(x)nh^{1+2j}} + o_p\left(\frac{1}{nh^{1+2j}}\right), &(13)\\ \mathrm{Bias}(\hat{m}^{(j)}(x,\alpha)|\mathbb{X}) &= e_{j+1}^T S^{-1}c_p\frac{j!}{(p+1)!}m^{(p+1)}(x)h^{p+1-j} + o_P(h^{p+1-j}). &(14) \end{aligned}$$

Both formulas are the same as those for local polynomial fitting (Fan & Gijbels, 1996, Theorem 3.1). Note that for symmetric kernels 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, 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 (11) into account:

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

$$\text{Bias}(\hat{m}^{(j)}(x,\alpha)|\mathbb{X}) = e_{j+1}^T S^{-1}c_p \frac{j!}{(p+1)!} m^{(p+1)}(x) h^{p+1-j} + o_P(h^{p+2-j}) \quad (15)$$

and for $p-j$ even

$$\text{Bias}(\hat{m}^{(j)}(x,\alpha)|\mathbb{X}) = \frac{e_{j+1}^T j!}{(p+1)!}\left[\left(\frac{\alpha'(x)}{\alpha(x)} + \frac{f'(x)}{f(x)}\right)\left(S^{-1}\tilde{c}_p - S^{-1}\tilde{S}S^{-1}c_p\right) m^{(p+1)}(x) + S^{-1}\tilde{c}_p \frac{m^{(p+2)}(x)}{p+2}\right] h^{p+2-j} + o_P(h^{p+2-j}). \quad (16)$$

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 (16) vanishes for

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

and this differential equation is solved for

$$\alpha_{opt}(x) \propto \frac{1}{f(x)}\,. \quad (17)$$

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 (16) and (17), 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 (17), which suggest to set $k = -1$ in (4), is that this seems to be in contrast to the proposal $k = 1$ 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

$$(\hat{m}(x_1), \ldots, \hat{m}(x_n))^T = Ly. \tag{18}$$

The leverage (or influence) values $\mathrm{infl}(x_i)$ are the diagonal elements $l_i(x_i)$ of L and measure the sensitivity of the fitted curve $\hat{m}(x_i)$ to the individual data points (Loader, 1999). As illustrated by Loader in Fig. 2.6, the leverage values of a local fit rise strongly near the boundary, implying that the boundary values, in particular horizontal outliers, have a huge impact on the fitted curve. Hence, a promising way to robustify against outlying predictors would be to control the leverage values in these regions. This would also implicitly control the variance, as the influence function $\mathrm{infl}(x)$ serves as an upper bound for the variance function $\mathrm{Var}(\hat{m}(x))$ (Loader, 1999, Theorem 2.3).

Let us consider for simplicity the manipulated Nadaraya-Watson estimator (6). The leverage values of this estimator are given by

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

where $\hat{f}_\alpha(x) = \sum_{j=1}^n K_h(x_j - x)\alpha(x_j)$ may be seen as a weighted kernel density estimator at point x, which corresponds to a usual kernel density estimator if $\alpha(x_j) \equiv 1/n$. From (19) we see that the leverage values are constant iff

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

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 left, 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 more stable for $k = 1$. For a local polynomial fit of second order these differences are not as pronounced, but the tendency is still observable (Fig. 2 right). The differences between the weighting schemes vanish for larger samples: Asymptotically, the influence function does not depend on $\alpha(\cdot)$; for instance one has $\mathrm{infl}(x) = K(0)/(nhf(x)) + o((nh)^{-1})$ for the local constant estimator (6).

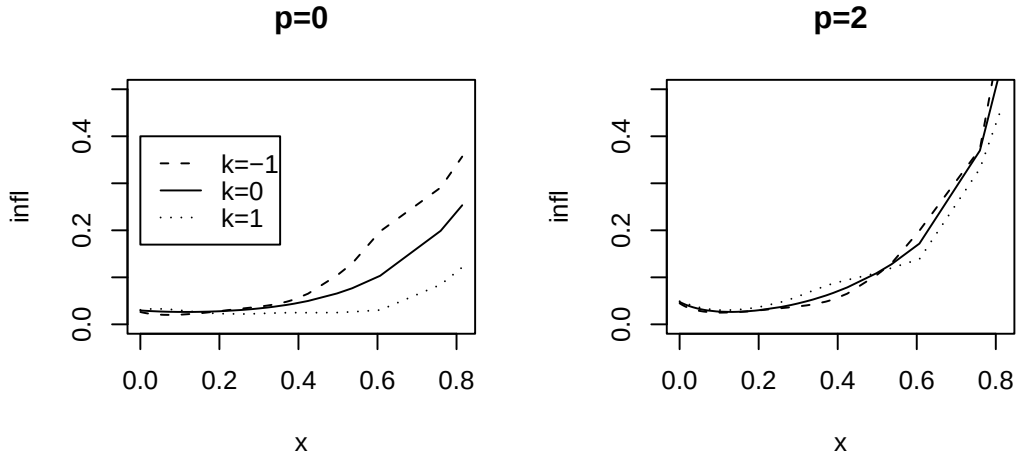

Figure 2: Leverage (influence) values for Beta-distributed design ($n = 50$) for different polynomial degrees and weighting schemes.

It should be noted that leverages values have attracted some attention previously 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 down-weighting high leverage points is the Mallows-estimator (Mallows, 1983). In the

literature on parametric robust regression, *leverage values* as defined in (18) are often used as an indicator for *leverage points*, i.e. outliers in the design space. However, as pointed out by Rousseeuw & van Zomeren (1990), leverage values may be very weak in detecting leverage points as they suffer from the masking effect. In order to avoid this semantical ambiguity, we do not use the term *leverage points* at all in this paper, but speak instead of outliers in the design space, or, equivalently, horizontal outliers or outlying predictors.

2.3 Behavior at the boundary

We have at this point two weighting schemes, which are both in some (but different!) sense optimal, or at least plausible. In order to get some insight into this apparent contradiction, we firstly observe from Fig. 2 that points associated with high leverage values 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. Let us therefore investigate the asymptotic properties of boundary points. We assume that the design density f has a bounded support (say, w.l.o.g, $[0,1]$) 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 $x = ch$*

$$\begin{aligned}
\mathrm{Var}(\hat{m}^{(j)}(x,\alpha)|\mathbb{X}) &= e_{j+1}^T S_c^{-1} S_c^* S_c^{-1} e_{j+1} \frac{(j!)^2 \sigma^2(0+)}{f(0+) n h^{1+2j}} + o_P\left(\frac{1}{nh^{1+2j}}\right), \\
\mathrm{Bias}(\hat{m}^{(j)}(x,\alpha)|\mathbb{X}) &= e_{j+1}^T S_c^{-1} c_{p,c} \frac{j!}{(p+1)!} m^{(p+1)}(0+) h^{p+1-j} + o_P(h^{p+1-j}),
\end{aligned}$$

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 (17) are no longer valid for a boundary point. Practically, this observation is not yet very useful. Every data set consists of interior and boundary points, but our weights $\alpha(x_i)$ do not depend on the location x. Hence, one has to choose *one* weight function that serves them all.

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 $\alpha \sim f$ was originally introduced to robustify against outlying predictors, which is, as one might argue, rather a finite sample problem, suggesting the simple rule: Use $\alpha \sim f$ for small sample sizes, and the asymptotically optimal weights $\alpha \sim 1/f$ for large sample sizes.

3.1 A tutorial on the effect of outlying predictors

To investigate this, we generate data from the underlying function $m_1(x) = \sqrt{x}$ and Beta-distributed design as in Section 2.2, for sample sizes $n = 50$ (Fig. 3 a,b) and $n = 1000$ (c,d). In (a) and (c) we have situations where either no relevant outlying predictors are present, or, if they are, their associated responses are distributed roughly symmetrically around m_1. In this case, the fits using $k = 1, 0, -1$ are very similar, except that the 'robust' weighting scheme $k = 1$ may give too much weight to the *previous* observations compared to the next ones, resulting in a slight oversteering. The matter is different to the right (b, d), where we observe that the asymptotic weighting scheme can produce a heavy bias in sparse data regions, whereas the robust version still stays comparatively near to the underlying function. Further, we observe that the problems with outlying predictors do not necessarily alleviate for increasing sample size n. Though the beginning of the γ_n-outlying region (vertical lines in Fig. 3) moves to the right for increasing n, there is a constant probability γ (here: $\gamma = 0.2$) of observing values beyond it, and these data points may have the same disturbing influence on the

fit as for smaller sample sizes. Hence, horizontal outliers cannot be considered as just a finite sample problem.

For comparison, we consider a second data set with interior sparse design (though one would not speak of *outlying* predictors in this case), generated from $m_2(x) = x + 2\exp(-x^2)$, with the design being an equal mixture of a Beta$(2, 9)$ and a Beta$(9, 2)$ distribution ($n = 50$, Fig. 3 e,f). Here, as expected by the theory, the asymptotic weights are rather superior. Concretely, if there is a single sparse design point as in (f), they are strongly better, while all fits are similar if the sparse design points are roughly symmetrically distributed around m_2, as in (e). We omit the plot for $n = 1000$ as it does not give new insight.

We see that there is no guarantee that either weight function improves the fit generally. From our first impression, the asymptotic weight will make better sense for interior than for boundary sparse design, in concordance with the theory. In the latter case, however, a statement seems to be impossible at this point, as the behavior of the fit depends on the position and number of outlying design points, and these may occur for any sample size and (bounded or unbounded) design. A simulation study is evidently called for, and we give the results below.

3.2 Simulation study

For data sets of size $n = 50$ and $n = 1000$, each 1000 replicates were generated as above, for both m_1 and m_2. 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 $\int \ell(\hat{m}(x) - m(x))\,dx$, 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 for instance Fahrmeir & Tutz (2001), p. 190, for an overview. We will work representatively with the three choices outlined above, ensuring that the results are not a particular feature of a certain error criterion. The results of the simulation study are shown in Fig. 4.

We start with the case that turned out to have the simpler interpretation,

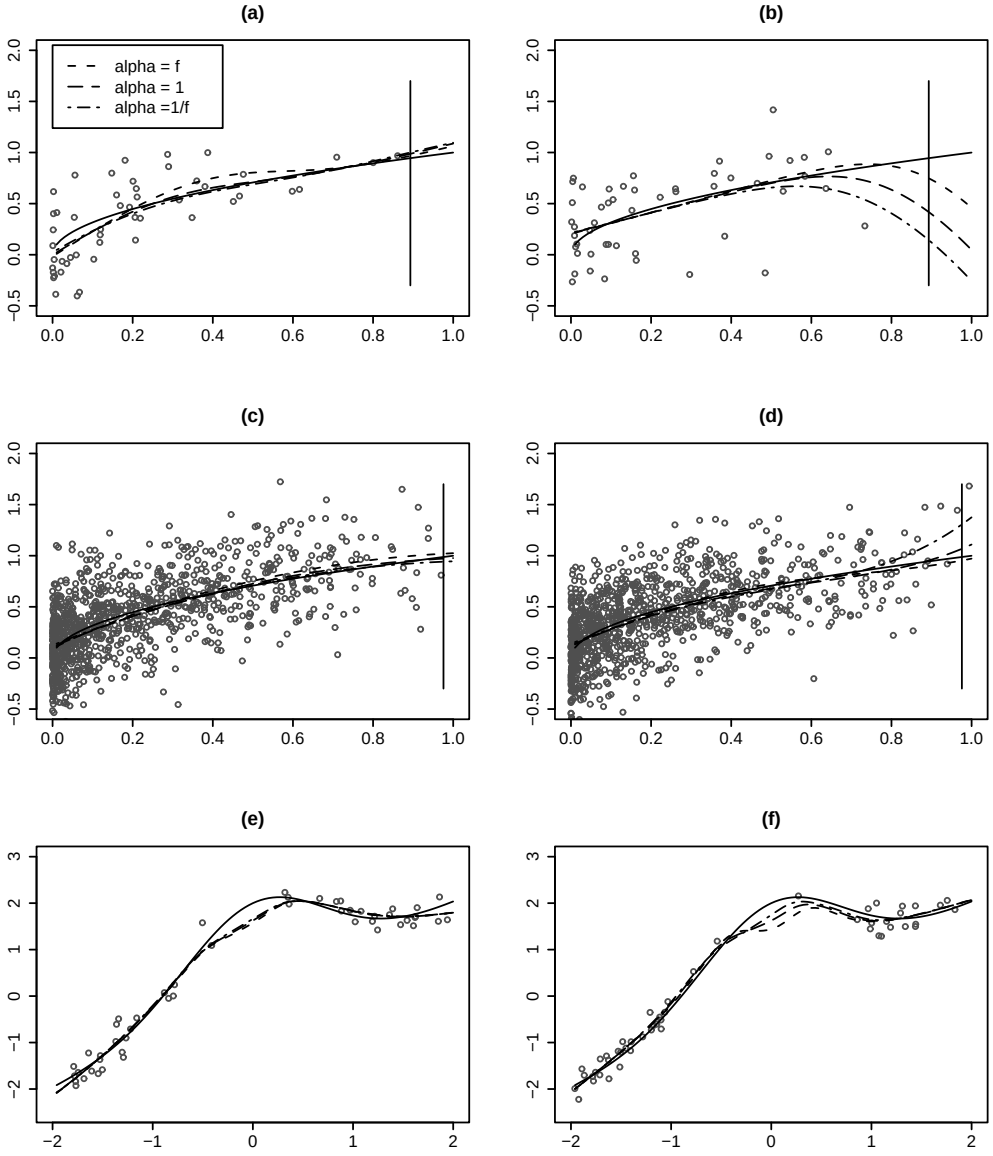

Figure 3: Selected examples: Design-weighted estimates for sample sizes $n = 50$ (a,b,e,f) and $n = 1000$ (c,d), underlying functions $m_1(x) = \sqrt{x}$ (a,b,c,d) and $m_2(x) = x + 2\exp(-x^2)$ (e,f), each estimated using weights $\alpha = \hat{f}$, $\alpha \equiv 1$ and $\alpha = 1/\hat{f}$, and polynomial order $p = 0$ (a,b,c,d) or $p = 2$ (e,f), respectively. The predictors follow a Beta$(0.5, 2)$ distribution in (a,b,c,d), and a mixture of a Beta$(2, 9)$ and Beta$(9, 2)$ distribution in (e,f). True functions are indicated by solid lines; vertical lines in (a,b,c,d) indicate the beginning of the γ_n- outlying region ($\gamma = 0.2$) at 0.893 and 0.976, respectively. Abscissa: x; ordinate: y.

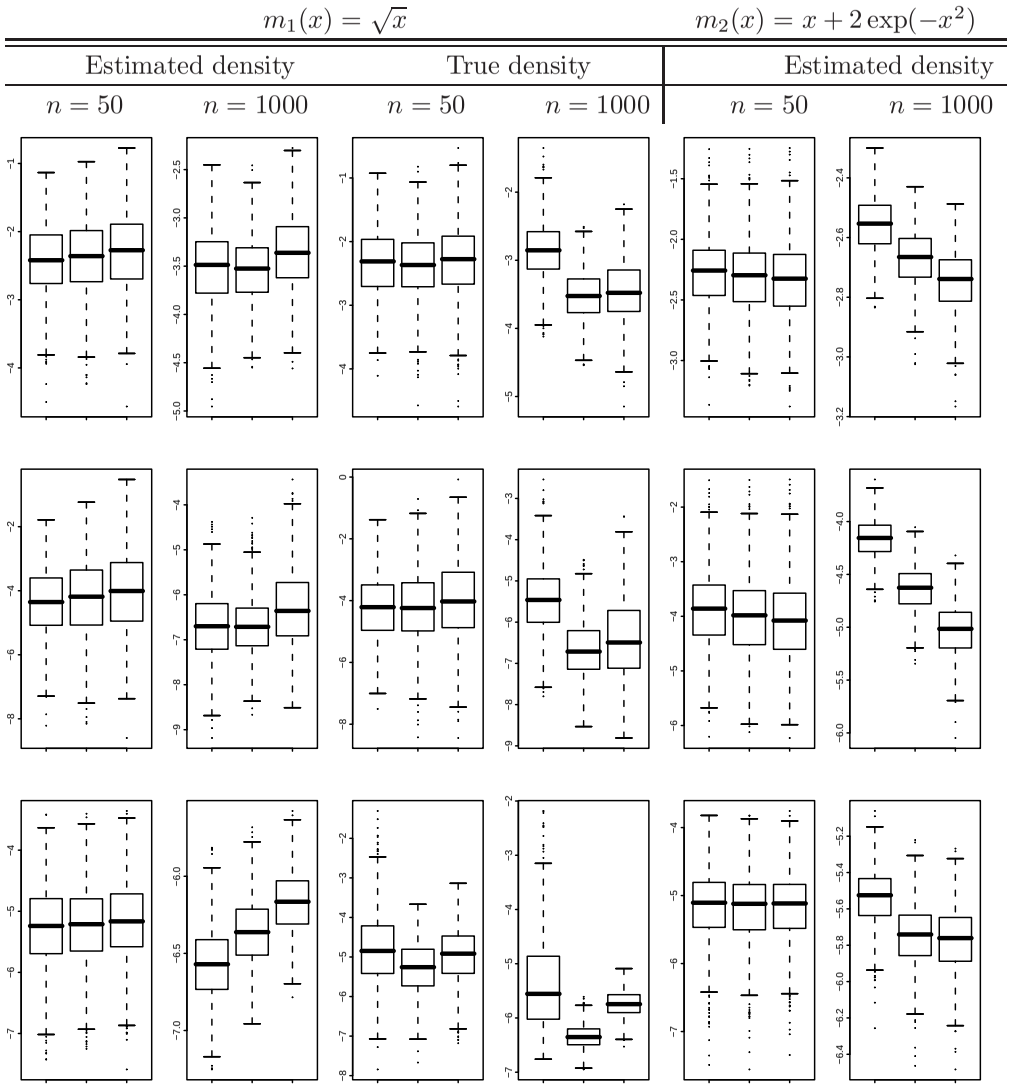

Figure 4: Design-weighted local regression: Boxplots of $\log(IAE)$ (top), $\log(ISE)$ (middle), and $\log(ASE)$ (bottom) over 1000 simulated data sets, each with weight functions $\alpha = f$, $\alpha = 1$, and $\alpha = 1/f$ (from left to right). Note that the figures have differing locations on the logarithmic scale, as not the absolute values are of interest, but rather the relative differences between weighting schemes.

which was expectedly the function m_2. In columns 5 and 6 of Fig. 4 the logarithms of the three error criteria are given for samples of size $n = 50$ and $n = 1000$, respectively. We apply the weighting schemes $\alpha = \hat{f}$, $\alpha \equiv 1$ and $\alpha = 1/\hat{f}$ (i.e. $k = 1, 0, -1$; from left to right within the boxplots), where f was estimated using the kernel density estimator $\hat{f}(x) = \frac{1}{ng} \sum_{i=1}^{n} K\left(\frac{x_i - x}{g}\right)$. The bandwidth g was selected for each simulated data set anew using Silverman's (1986, p. 48) bandwidth selector, as also proposed in Einbeck, André & Singer (2004). The asymptotic weights (the third boxplot each) turned out to be superior for both sample sizes and all three error criteria. This superiority was larger for $n = 1000$ than for $n = 50$, which was to be expected for an asymptotically derived rule. We note that the superiority of the asymptotic weights decreases when the Gaussian noise ($\sigma = 0.2$) is increased (not shown), but the situation does not change substantially.

We turn our attention now to m_1, contaminated with Gaussian noise ($\sigma = 0.3$), and carry out the same study as above. From the first column in Fig. 4 we see that, for all error criteria, there seems to be some tendency that the robust weights $\alpha = \hat{f}$ are superior. For a larger sample size, the asymptotic weights produce even worse results, 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. This is even more remarkable as we did not assume 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 to which we referred in the introduction is actually of size $n = 1067$, giving another example that the usefulness of robust weighting is not restricted to small sample sizes.

It is conceivable that the performance of the asymptotic weights depends on the accuracy of the density estimate. Therefore, we repeated the analysis using the true density, and, indeed, the asymptotic weights perform much butter for either sample size (Fig. 4, columns 3 and 4), though they never succeed to be the 'winning' weight. We return to this observation in the next section after our look at sampling theory.

We have to stress that the picture might be different in other situations. We

did 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 sometimes on the asymptotic side. (We note that these differences are not attributable to insufficient Monte Carlo precision of the simulation study; for a specific underlying function, the results are repeatable, and the differences between weighting schemes are mostly strongly significant.) The general pattern concerning our initial hypothesis, however, was mostly the same: The asymptotic weights *may* perform very well compared to their competitors, particularly for interior sparse design, and *if they perform well*, then their performance improves with the sample size. However, they behave extraordinarily hazardous, and they still might give very poor results for large sample sizes, as their success depends dramatically on the existence and position of outlying predictors, and also on the accuracy of the density estimate.

We do not seem to have reached the bottom of the barrel yet. 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 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. 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 match the proportions in the population, keeping the bias small requires weighting the strata accordingly.

4.1 Design-weighted local smoothing and the Horvitz-Thompson estimator

There has been considerable efforts in the sampling literature in the search for some theoretical grounding or justification for weighting a sample. One of the most influential theoretical results in this connection was established by Horvitz & Thompson (1952). From a population $Y_1, \ldots, Y_N$ we draw without replacement a sample of size n. Suppose the population total $Y = \sum_{i=1}^{N} Y_i$ is to be estimated. We define the random variable δ_i by $\delta_i = 1$, if unit i is found in the sample, and 0 otherwise. Horvitz and Thompson (hereafter: HT) showed that among all linear estimators of the form $\hat{Y} = \sum_{i=1}^{N} \alpha_i \delta_i Y_i$ the estimator

$$\hat{Y}_{HT} = \sum_{i=1}^{N} \delta_i \frac{Y_i}{\pi_i}, \tag{21}$$

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. 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 from the ℓ-th stratum is given by

$$\pi_\ell = \frac{np_\ell}{NP_\ell} \tag{22}$$

where p_ℓ and P_ℓ are the proportions of the $\ell-$th stratum in the sample and in the population, respectively (see e.g. Kish, 1965, p. 92). DuMouchel & Duncan (1983) linked this concept to parametric regression by applying weights inversely proportional to (22) in a minimization problem of type (2) in the special case $h \longrightarrow \infty$.

For the interpretation of these results, recall that (17) means that the bias is minimized when the observations are weighted with the inverse density, while HT 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 (17) as an asymptotic and nonparametric version of HT'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	$\alpha_\ell = 1/\pi_\ell \sim P_\ell/p_\ell$	Adaptation 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* into 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 legitimacy is much less acknowledged (e.g., 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.

This brings us back to the problem discussed in the previous section. When replacing the true design density f with an estimated one, $\hat{f}$, the asymptotic results do not apply either, and the asymptotic weights (17) are no longer 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 (20) was explicitly based on the *estimated* density. Hence, it is no contradiction 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.

4.2 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 probability 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

$$\pi_i = nY_i/Y, \tag{23}$$

the estimator $\hat{Y}$ has zero variance and mean square error and the sampling will be optimal. Obviously, the design used in the fable is far from optimality in the sense of (23); it is rather 'about as poor a design imaginable' (Overton & Stehman, 1996). 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).

4.3 A general dilemma?

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 (17), one easily verifies that the variance term (12) vanishes if $f(x) \propto \sigma^2(x)$, which is then the analogous formula to (23) and leads to weights proportional to (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 generally: Aiming to minimize the bias, statistical theory suggests that weights be chosen 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.

5 Conclusion

We have so far studied the properties of design-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 straightforward idea to rely on the sample size turned out to be rather misleading. It seems to play some role whether the design density is known or estimated. However, for our example function m_1 the asymptotic weights could not compete with the simple constant weights even when using the true design density. Furthermore, it is beyond common sense to base the choice of the weight function not on the design 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. As a consequence, a more subjective viewpoint is helpful. The asymptotical result (17) confirms the statement by Hastie & Loader (1993), who called an endpoint *'the most informative observation'* when fitting at this endpoint. Einbeck, André & Singer (2004) added that this holds only when this point can be considered as *'as reliable as in the interior'*. 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, went in a similar direction when considering, in the parametric setting, *'extreme design points (which might be wrong)'*. 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 in doubt stay with the usual constant weights (i.e. $k = 0$). Though the asymptotically optimal weights ($k = -1$) have a clear potential to improve the fit in special situations, they 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 with 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 short since it follows mainly the lines of the corresponding proof for local polynomial fitting, see Fan, Gijbels, Hu & Huang (1996). Let

$w_i = K_h(x_i - x)$ and

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

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

Bias: Using standard asymptotics reveals that

$$r_{n,j} = nh^j (f_\alpha(x)\mu_j + h f'_\alpha(x)\mu_{j+1} + o_n), \tag{24}$$

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

$$R_n = nH[f_\alpha(x) S + h f'_\alpha(x)\tilde{S} + o_n]H \tag{25}$$

holds. Then, using Taylor's expansion and equation (7), we get

$$\mathrm{Bias}(\hat{\beta}|\mathbb{X}) = R_n^{-1}\left[\beta_{p+1} d_n + \beta_{p+2}\tilde{d_n} + o_P(\tilde{d_n})\right], \tag{26}$$

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

$$R_n^{-1} = \frac{1}{n} H^{-1}\left[\frac{1}{f_\alpha(x)} S^{-1} - h\frac{f'_\alpha(x)}{f^2_\alpha(x)} S^{-1}\tilde{S}S^{-1} + o_n\right] H^{-1}. \tag{27}$$

Plugging (27) into (26), and substituting (24) into the vectors d_n and $\tilde{d_n}$, yields (9) via some simple matrix algebra, taking into account that $f'_\alpha(x)/f_\alpha(x) = \alpha'(x)/\alpha(x) + f'(x)/f(x)$.

Variance: Similar like (25) we find that

$$R^*_n = \frac{n}{h} H[s_\alpha(x) S^* + h s'_\alpha(x)\tilde{S}^* + o_n]H, \tag{28}$$

where $s_\alpha(x) = \sigma^2(x)\alpha^2(x) f(x)$. Substituting (28) and (27) in $\mathrm{Var}(\hat{\beta}|\mathbb{X}) = R_n^{-1} R^*_n R_n^{-1}$, we derive (10) using 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.)*. Chichester: 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.

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., 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. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models (2nd Ed.)*. 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. *Stat. Sin.* **6**, 113–127.

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

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". *Stat. Sci.* **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.

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. (1983). Robust methods. In R. Gnanadesikan (Ed.), *Proceedings of Symposia in Applied Mathematics*, pp. 49–74. Providence, RI: Amer. Math. Soc.

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

Overton, W. S. and Stehman, S. V. (1996). Reply to: D.V. Lindley: Letter to the editor. *Amer. Statist.* **50**, 197–198.

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

Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *J. Amer. Statist. Assoc.* **85**, 633–639.

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.

Department of Mathematical Sciences, Durham University, South Road, Science Laboratories, Durham DH1 3LE, UK, Tel +44 191 3343125, Fax +44 191 3343051

E-mail: jochen.einbeck@durham.ac.uk

Department of Statistics, University of Munich (LMU), Ludwigstr. 33, 80539 Munich, Germany

E-mail: augustin@stat.uni-muenchen.de