Smoothing, Sampling, and Basu's elephants

Jochen Einbeck

jochen.einbeck@durham.ac.uk

Milton Keynes, 24th of April 2009

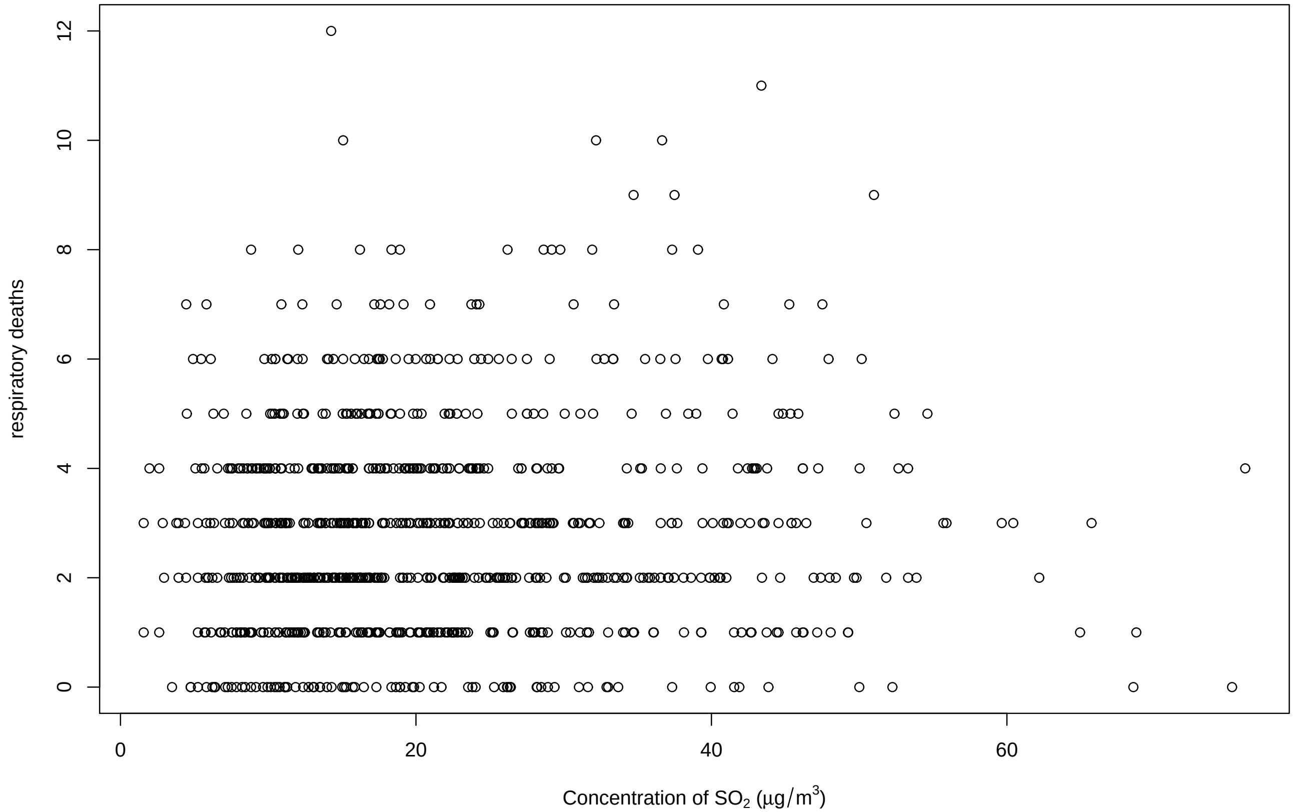
joint work with Thomas Augustin, LMU München, and Julio M. Singer, Universidade de São Paulo

Motivation

A bright and sunny day in São Paulo....

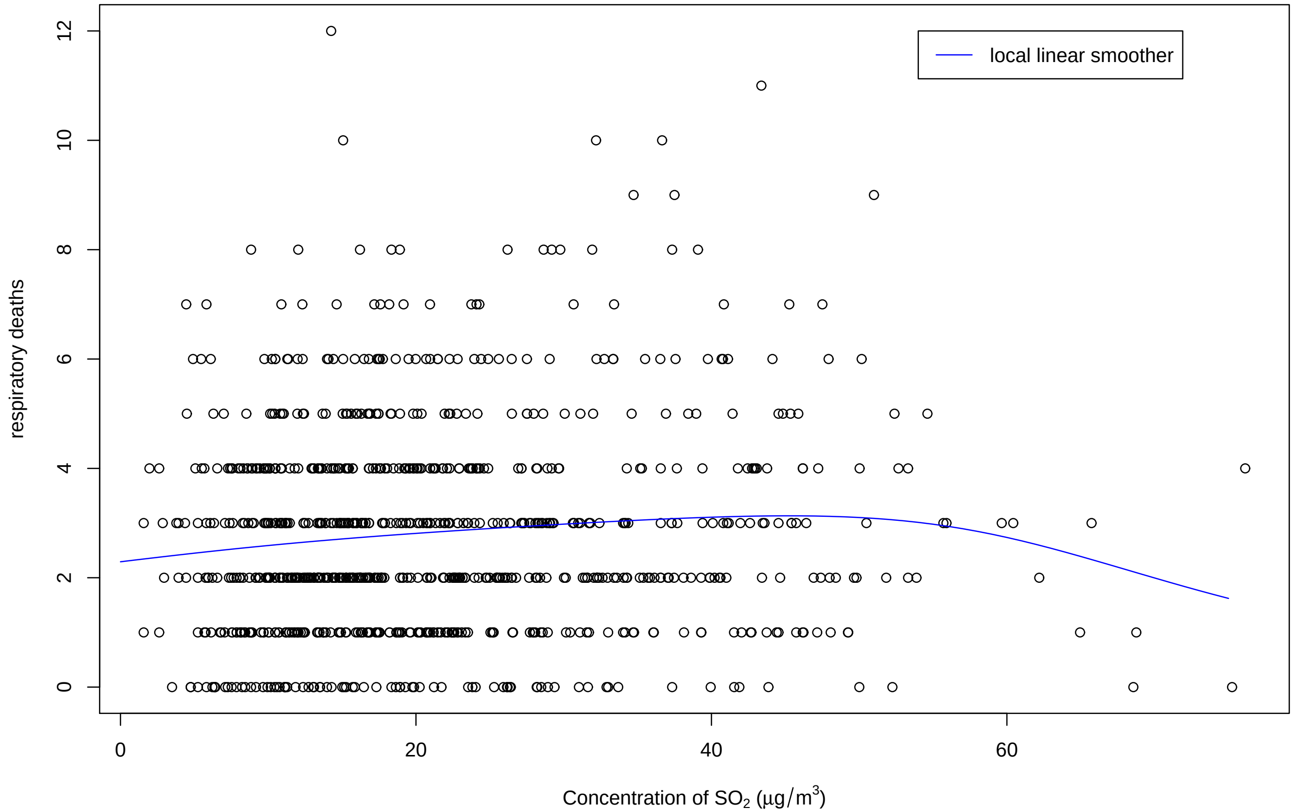
Motivation (cont.)

- Data: Respiratory deaths of children under five in the city of São Paulo, 1994–1997.



Motivation (cont.)

- Data: Respiratory deaths of children under five in the city of São Paulo, 1994–1997 with nonparametric smoother.

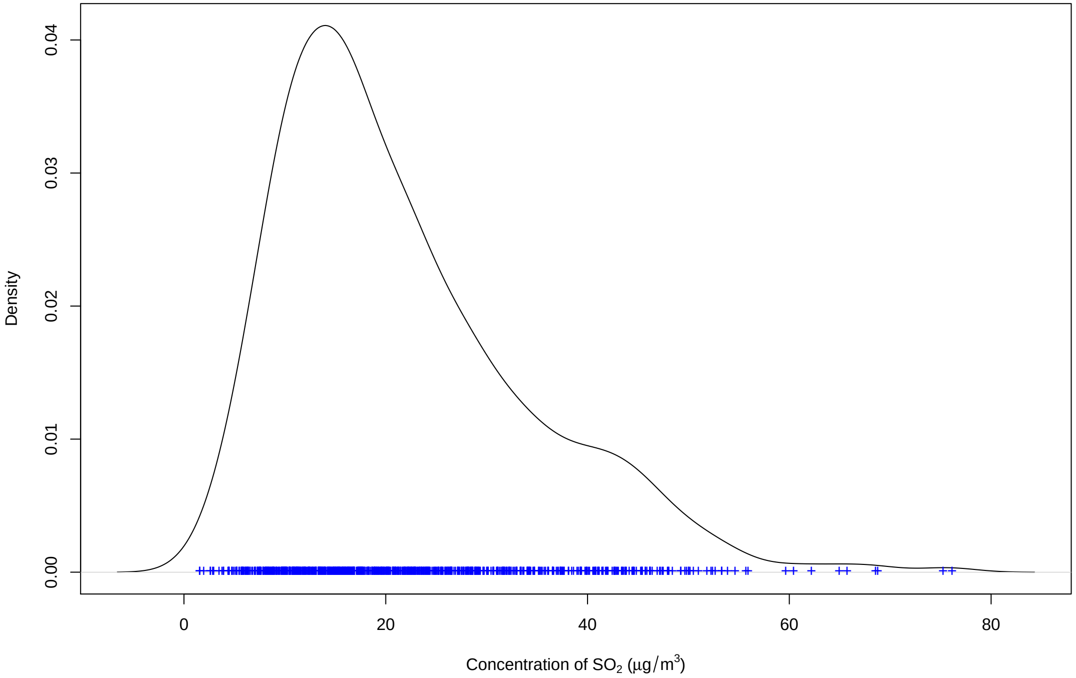


Motivation (cont.)

- One observes that the effect of the pollutant on the infant mortality decreases for high pollutant concentration – a result which has no biological plausibility.
- The smooth curve is pulled down by two observations near the right boundary. Hence, the problem is due to outliers in the design space.
- Horizontal outliers have attracted far less attention in the (nonparametric) statistical literature than vertical outliers.
- Nonparametric smoothers robustifying against outlying response will not necessarily robustify against outlying predictors!

Motivation (cont.)

- Horizontal outliers correspond to sparse boundary regions, i.e. to regions close to the boundary with small design density:



- These regions with sparse design density may give unreliable information, and this even when the data associated to those regions are not outlying in $y-$ direction.

Design-weighted local smoothing

- The idea is to use the estimated design density as a weight function in the local linear regression problem.
- This will reduce the influence of outliers in the design space.
- Concretely, let $m(\cdot)$ denote the true underlying function, then a design-weighted local smoother is obtained as $\hat{m}(x) = \hat{\beta}_0(x)$, where $\hat{\beta}_0(x)$ and $\hat{\beta}_1(x)$ minimize

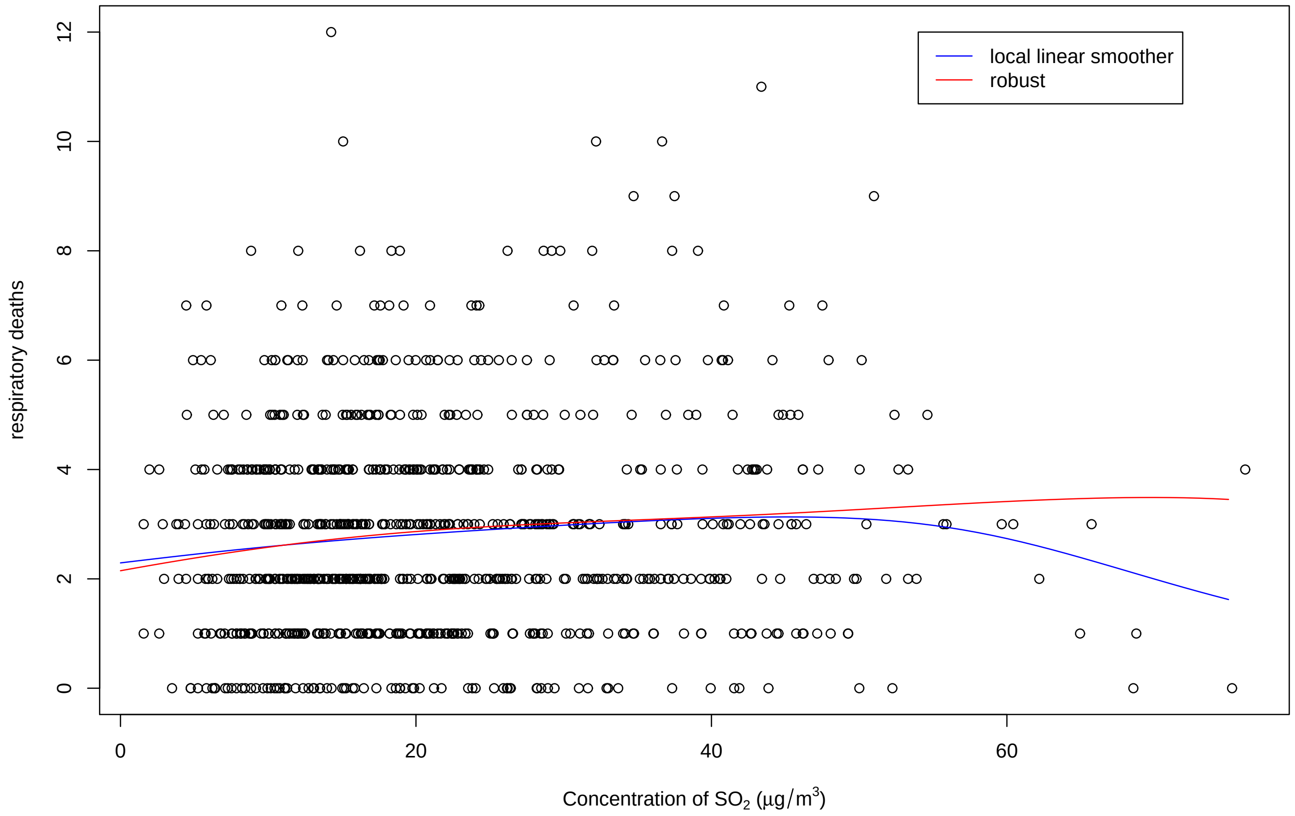
$$\sum_{i=1}^{n} (Y_i - \beta_0(x) - \beta_1(x)(x - X_i))^2 \, \alpha(X_i) K\left(\frac{x - X_i}{h}\right) \qquad (1)$$

  with respect to $\beta_0(x)$ and $\beta_1(x)$.
- In order to robustify against outliers in the design space, we choose

$$\alpha(x) = \hat{f}(x) = \frac{1}{ng} \sum_{i=1}^{n} K\left(\frac{X_i - x}{g}\right).$$

Design-weighted local smoothing (cont.)

- Robustified curve in the respiratory data example (*Einbeck, André, & Singer, Environmetrics, 2004*):



- (Alternative: Smoothing with Monotonicity constraint: *Leitenstorfer & Tutz, Biostatistics, 2007*)

Design-weighted local smoothing (cont.)

- Next, we investigate theoretically a generalized version of (1), namely the design-weighted least squares problem

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

  with some general, continuously differentiable function α.

- From the vector $(\hat{\beta}_0(x), \ldots, \hat{\beta}_p(x))$ minimizing (2), one gets estimators of
  - the regression function m: $\hat{m}(x) = \hat{\beta}_0(x)$.
  - its derivatives $m^{(j)}, j = 1, \ldots, p$: $\hat{m}^{(j)}(x) = j!\hat{\beta}_j(x)$.
- Note that there are two kinds of weights involved here
  - *kernel weights* $K(\cdot)$ (depend on *distance* of x_i and x)
  - *design weights* $\alpha(\cdot)$ (depend only on *location* of x_i)

Asymptotics

- **Theorem.** *Let* $h \longrightarrow 0$ *and* $nh^3 \longrightarrow \infty$, *and* $\mathbb{X} = (x_1, \dots x_n)$. *Under regularity assumptions we get for* $p - j$ *odd*

$$\mathrm{Bias}(\hat{m}^{(j)}(x)|\mathbb{X}) = e_{j+1}^T S^{-1} c_p \frac{j!}{(p+1)!} m^{(p+1)}(x) h^{p+1-j} + o_P(h^{p+2-j}) \tag{3}$$

and for $p - j$ *even*

$$\begin{aligned}\mathrm{Bias}(\hat{m}^{(j)}(x)|\mathbb{X}) \quad &= \quad e_{j+1}^T \frac{j!}{(p+1)!} \left[\left(\frac{\alpha'(x)}{\alpha(x)} + \frac{f'(x)}{f(x)} \right) s_p m^{(p+1)}(x) + \right. \\ &\left. + S^{-1} \tilde{c}_p \frac{m^{(p+2)}(x)}{p+2} \right] h^{p+2-j} + o_P(h^{p+2-j}). \end{aligned} \tag{4}$$

where $s_p = (S^{-1}\tilde{c}_p - S^{-1}\tilde{S}S^{-1}c_p)$, S, $\tilde{S}$, c_p, $\tilde{c}_p$ *are matrix/vector- valued constants consisting of kernel moments.*

- Interestingly, (3) is the same as for usual (unweighted) local polynomial smoothing, while (4) is not.

Asymptotics (cont.)

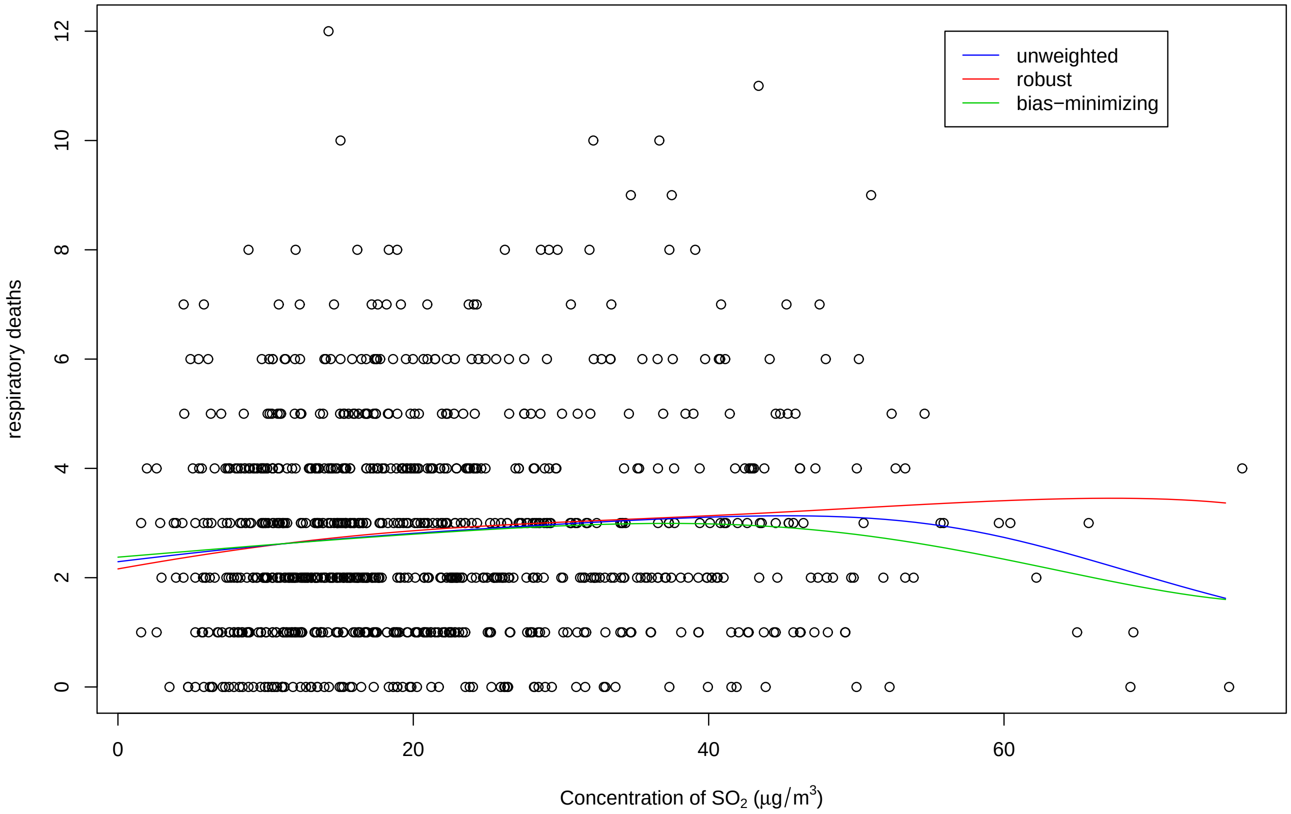
- The more interesting of the two expressions above is the second one, because it shows that in this case the leading term is *not* independent of $\alpha(\cdot)$. This gives the chance to reduce the bias.
- The first term in the squared bracket in (4) vanishes for $\alpha'(x)/\alpha(x) + f'(x)/f(x) = 0$,
- This differential equation is solved for

  $$\alpha_{opt}(x) = c\frac{1}{f(x)}, \tag{5}$$

  with $c \in \mathbb{R} \setminus \{0\}$.
- This seems to be in conflict with the “robust” weights suggested beforehand!

Comparison

- Fit a local quadratic smoother ($p = 2$) to respiratory data:



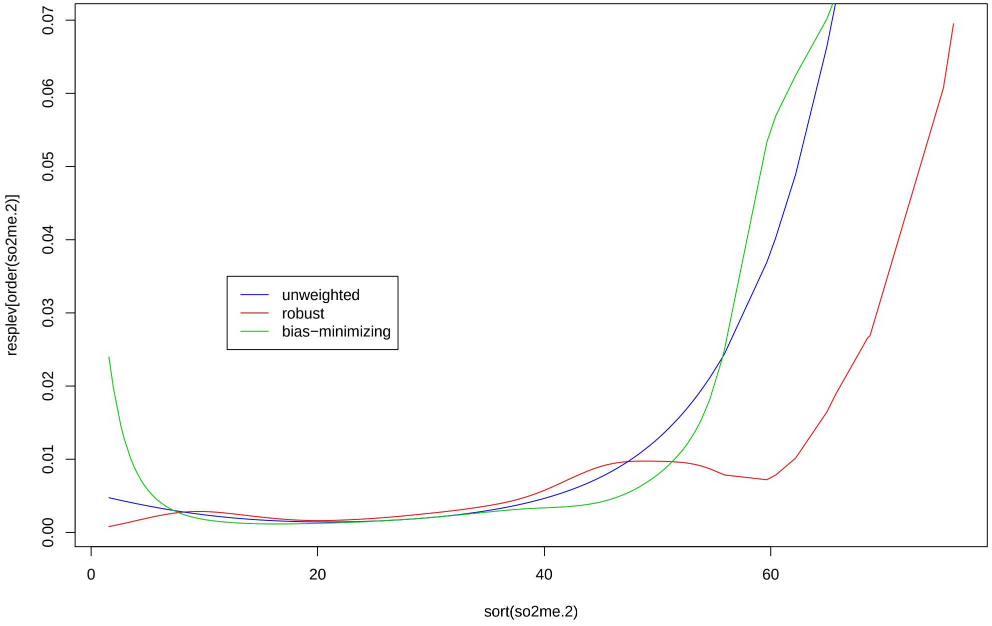
- Apparently, the bias-minimizing smoother makes things rather worse.

Leverages

- The hat matrix S (“smoothing matrix”) of a smoother $\hat{m}$ is defined as

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

- The leverage values values are the diagonal elements $s_i(x_i)$ of S and measure the sensitivity of the fitted curve $\hat{m}(x_i)$ to the i—th data point.
- Leverage values for respiratory data with local quadratic smoothers:



Dilemma?

- The weights $\alpha \sim \frac{1}{f}$ reduce the bias.
- At the same time, they increase the leverages near the boundary, and therefore the sensibility of the fitted curve to outliers in the predictor space.
- On the other hand, the robust weights $\alpha \sim f$ reduce the variance of the fit in boundary regions:

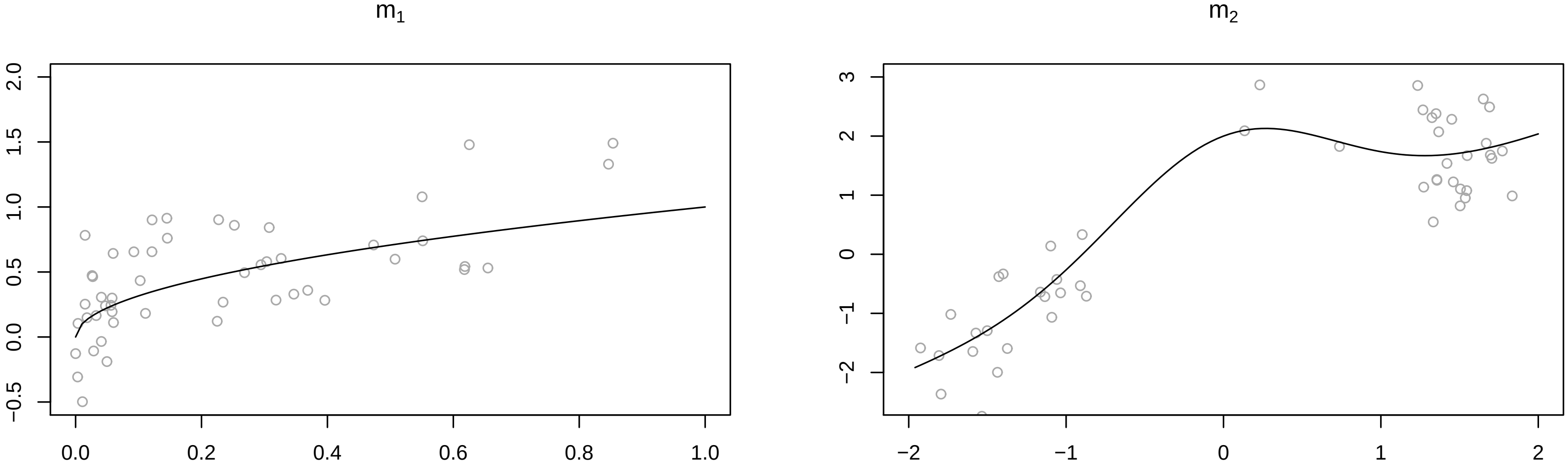
$$\mathsf{Var}(\hat{m}(x_i)) \leq s_i(x_i)$$

  (Loader, 1999).
- So, which to choose?

Simulation study

- We consider each 1000 replicates of functions
  - $m_1(x) = \sqrt{x}$, $x \sim$ Beta$(0.5, 2)$, $y \sim N(0, 0.3^2)$.
  - $m_2(x) = x + 2\exp(-x^2)$,
    $x \sim 0.5 \cdot$ Beta$(2, 9) + 0.5 \cdot$ Beta$(9, 2)$, $y \sim N(0, 0.2^2)$.

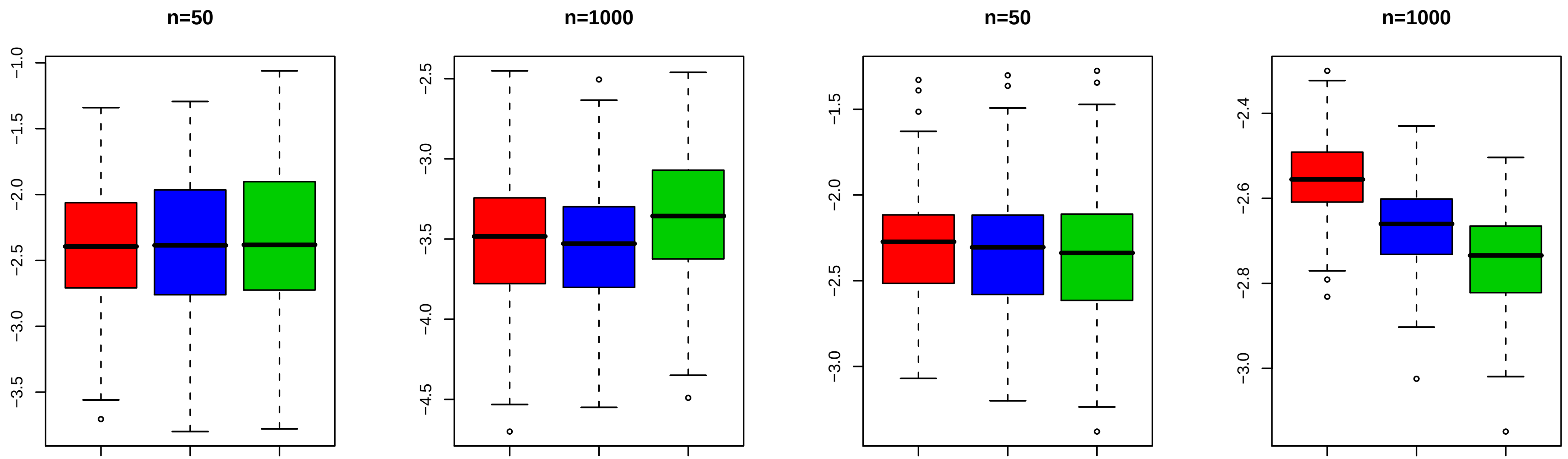
  and fit smoothers with weights $\alpha \sim \hat{f}$, $\alpha \sim 1$, and $\alpha \sim 1/\hat{f}$, to each replicate.



Simulation study (cont.)

- Each plot gives the logarithms of the integrated average errors $IAE = \int |\hat{m}(x) - m(x)|\, dx$ using the weights $\hat{f}$, 1, $1/\hat{f}$.

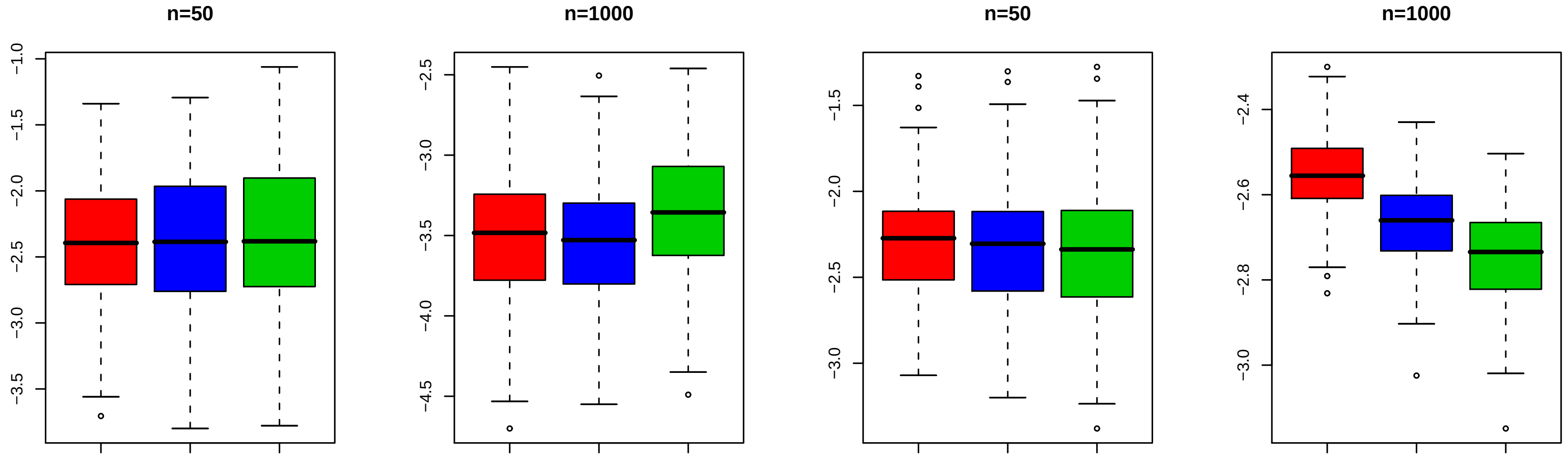
$m_1(x) = \sqrt{x}$ $\qquad\qquad$ $m_2(x) = x + 2\exp(-x^2)$



Simulation study (cont.)

- Each plot gives the logarithms of the integrated average errors $IAE = \int |\hat{m}(x) - m(x)|\, dx$ using the weights $\hat{f}$, 1, $1/\hat{f}$.

$m_1(x) = \sqrt{x}$ $\qquad\qquad$ $m_2(x) = x + 2\exp(-x^2)$



- Results do not suggest that “Selecting weights depending on the sample size” is generally a useful strategy.

The link to sampling theory

- Design-weighting is very common and has been intensively discussed in *sampling theory*.
- One of the most important theoretical justifications for design-weighting was given by Horvitz & Thompson (HT, 1952):
- For a sample of size n drawn from a population $Y_1, \dots, Y_N$, they showed that among all linear estimators of the form $\hat{Y} = \sum_{i=1}^{N} \alpha_i \delta_i Y_i$, the HT estimator

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

is the only unbiased estimator for the population total, Y.

(where π_i is the probability that the i-th element is drawn in any of the n draws and δ_i is an indicator taking the value 1 if unit i is selected)

The link to sampling theory (cont.)

- Note that this is a very similar result to that one obtained for design-weighted local smoothing:
- In both cases, one uses a bias-minimizing criterion and derives optimal weights inversely proportional to the selection probability / design density.
- More clearly:

Estimator	Bias minimized for	Interpretation
Horvitz-Thompson	$\alpha_i = 1/\pi_i$	π_i = selection probability of unit i
local smoothing, p even	$\alpha(x_i) \sim 1/f(x_i)$	$f(x_i)$ = design density at point x_i

- Similar conflict for HT weights??

Basu’s elephants

A circus owner plans to ship 50 adult elephants and therefore needs a rough estimate of their total weight. As weighing elephants is quite cumbersome, he intends to weigh only one elephant and to multiply the result with 50. However, the circus statistician insists in setting up a proper sampling plan, and to use the Horvitz-Thompson estimator. They agree to assign 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. The probability for all other elephants is $1/4900$, including ‘Jumbo’, the biggest elephant in the herd. Naturally, Samba is selected, and the statistician estimates the total weight of the herd by $100/99$ times Samba’s weight according to Horvitz-Thompson. If Jumbo were selected, his large weight would even have to be multiplied by 4900 to get the ‘best linear unbiased estimator’ of the total weight! Certainly, after having given these advices, the circus statistician was sacked.

Basu’s elephants (cont.)

- Basu’s fable provoked an at least 20-year-long discussion in the statistical literature on the general applicability of the HT estimator.
- Where is actually the problem with Basu’s fable? HT state that if $\pi_i = nY_i/Y$, the estimator $\hat{Y}$ has zero variance and the sampling will be optimal.
- Obviously, the design used in the fable is far from optimality in this sense. It is rather ‘about as poor a design imaginable’ (Overton & Stehman, 1996).
- Though HT’s estimator can reduce the bias of an estimate *given* the inclusion probabilities, it may produce useless estimates if they are unfortunately chosen, i.e. if the π_i are unrelated to the Y_i’ (Rao, 1999).
- The HT estimator is still useful e.g. when a second variable X_i is used to construct π_i which are correlated to the Y_i (ratio estimation).

Basu's elephants (cont.)

- Summarizing, Basu's paradox is solved: One has to get the probabilities right.
- However, Basu denied vehemently 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.
- Instead, he gives the responsibility for the useless result entirely to the HT 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's refusal to adjust the π_i touches the problem that we have in the smoothing context, with the "outliers in the design space" corresponding to "Jumbo".

Optimal design?

- In the sampling context, the key to apply the HT estimator successfully was to use design weights π_i which are related to the values $Y_i, i = 1, \ldots, N$.
- Can we find an analogous criterion for the smoothing context? In other word, is there an optimal design density?
- Look at variances:

Theorem. *Let $h \longrightarrow 0$ and $nh \longrightarrow \infty$. Under regularity assumptions one gets*

$$\mathrm{Var}(\hat{m}^{(j)}(x,\alpha)|\mathbb{X}) = e_{j+1}^T \left[S^{-1}S^*S^{-1} + hV_\alpha^*(x)\right] e_{j+1} \frac{(j!)^2\sigma^2(x)}{f(x)nh^{1+2j}} + o_p\left(\frac{1}{nh^{1+2j}}\right), \quad (6)$$

where $S, \tilde{S}$ and S^* are constant matrices containing kernel moments, and

$$\begin{aligned} V_\alpha^*(x) &= \left(2\frac{\sigma'(x)}{\sigma(x)} + 2\frac{\alpha'(x)}{\alpha(x)} + \frac{f'(x)}{f(x)}\right) S^{-1}\tilde{S}^*S^{-1} \\ &- \left(\frac{\alpha'(x)}{\alpha(x)} + \frac{f'(x)}{f(x)}\right) \cdot \left(S^{-1}\tilde{S}S^{-1}S^*S^{-1} + S^{-1}S^*S^{-1}\tilde{S}S^{-1}\right). \end{aligned} \quad (7)$$

Optimal design? (cont.)

- The first-order variance term is independent of α.
- Using the bias-minimizing weights $\alpha = 1/f$ in the second-order term $V_\alpha^*(x)$, one observes that $V_\alpha^*(x)$ is minimized when

$$f(x) \propto \sigma^2(x). \tag{8}$$

- This is the analogon to the optimal HT selection probabilities $\pi_i = nY_i/Y$.
- It is questionable whether (8) is a realistic condition in practice: The density in smoothing problems is in the very most cases inherent to the data or driven by practical considerations.
- Note that, trivially, when $\sigma^2(x) \equiv \sigma^2$, then (8) tells us that optimally $f(x) \propto 1$ and hence $\alpha_{opt}(x) \propto 1$.

Conclusion

- For local smoothing, there seems to be no such thing as an objective criterion for design weight selection!
- If there is a good reason to distrust a particular region of the design space, either because the observations themselves are unreliable in some sense, or because there are simply very few observations, the robust weights $\alpha \propto f$ (or $\alpha \propto f^k, k \geq 1$) are a reasonable choice.
- The asymptotic weights $\alpha \propto 1/f$ *may* perform 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 design points, the shape of the design density (and on the accuracy of the density estimate, if unknown).
- In doubt, better do not use weights at all!

Conclusion (cont.)

- There exists an striking analogy between the theories of sampling and smoothing, leading to similar theoretical results and practical pitfalls.
- This could tell a more general lesson. Weighting is performed in many statistical disciplines. A usual way of motivating such weights is by theoretical, bias-minimizing criteria, which will often suggest choosing weights inversely proportional to some kind of selection probability (density). This makes the estimator very sensitive to extreme observations (corresponding to Jumbo in Basu's fable and the outlying predictors in the smoothing context).
- Hence, we advise to be careful with bias-minimizing estimators if there are any observations which might be labelled by the terms "extreme", "undesired", "outlying", "weak", and the like, and it is likely that this holds beyond the scope of sampling and smoothing.

Literature

Brewer (2002). Combined survey sampling inference. Arnold, NY [The elephant pic!].

Einbeck, André & Singer (2004). Local smoothing with robustness against outlying predictors. *Environmetrics* 15, 541–554.

Einbeck & Augustin (2009). On design-weighted local fitting and its relation to the Horvitz-Thompson estimator. *Statistica Sinica* 19, 103–123.

Einbeck, Augustin & Singer (2007). Smoothing, sampling, and Basu's elephants. *Proceedings of the 22nd International Workshop on Statistical Modelling*, Barcelona, 2007, pp. 245–248.

Loader (1999). *Local regression and likelihood.* New York: Springer.

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

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