Smoothing, Sampling, and Basu's elephants

Jochen Einbeck[1], Thomas Augustin[2] and Julio M. Singer[3]

[1] Department of Mathematical Sciences, Durham University (UK), `jochen.einbeck@durham.ac.uk`,
[2] Institut für Statistik, LMU München (Germany), `augustin@stat.uni-muenchen.de`,
[3] Instituto de Matemática e Estatística, Universidade de São Paulo (Brazil), `jmsinger@ime.usp.br`.

Abstract

We investigate design-weighted local smoothing and show that the optimal (bias-minimizing) weights have similar form and interpretation as the optimal weights given by the Horvitz-Thompson theorem known from sampling theory. We set forth that the hazards in using bias-minimizing weights apply to kernel smoothing, too, suggesting to be cautious with the application of bias-minimizing weights in general.

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.

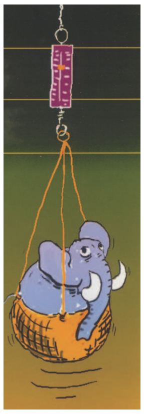

Fig. 1: Picture from cover of Brewer (2002)

Sampling and Weighting: The HT estimator

- This short version of a fable told by Basu (1971) illustrates his reservations against the Horvitz-Thompson (HT) estimator: For a sample of size n drawn from a population $Y_1, \ldots, Y_N$, Horvitz and Thompson (1952) showed that among all linear estimators of the form $\hat{Y} = \sum_{i=1}^N \alpha_i \delta_i Y_i$, the HT estimator
$$\hat{Y}_{HT} = \sum_{i=1}^N \delta_i Y_i / \pi_i$$
(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) is the only unbiased estimator for the population total, Y.

- Where is actually the problem with Basu's fable? Horvitz and Thompson state that if $\pi_i = nY_i/Y$, the estimator $\hat{Y}_{HT}$ has zero variance and sampling will be optimal. Rao (1999) warns that the HT estimator 'can lead to absurd results if the π_i are unrelated to the Y_i', and obviously the probabilities in the fable are far from optimal in this sense. Though HT's theorem can be used to reduce the bias of an estimate *given* the inclusion probabilities, it may produce useless estimates if they are unfortunately chosen.

- Nevertheless, HT's estimator proves to be 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 the known elephant weights X_i from the previous census, and to set $\pi_i = nX_i/X$, where X is the total weight of the herd measured at that time (Koop, 1971, in the discussion of Basu's essay).

Design-weighted local "bias-reducing" smoothing

Given a sample $(x_1, y_1), \ldots, (x_n, y_n)$ drawn from a bivariate population $(X, Y) \in \mathbb{R}^2$ with mean function $m(x) = E(Y|X = x)$, we are interested in a smooth estimate $\hat{m}(\cdot)$ of $m(\cdot)$. There are two forms of weighting that have to be distinguished here. Firstly, there are the *kernel weights* $K((x_i - x)/h)$, with a bandwidth h, and secondly, one can use additional *design weights*, $\alpha(\cdot)$, leading to the design-weighted least squares problem

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

From the vector $(\hat{\beta}_0(x), \ldots, \hat{\beta}_p(x))$ minimizing (1), one easily gets design-weighted local polynomial estimators of m and its derivatives, $\hat{m}^{(j)}(x) = j!\hat{\beta}_j(x)$.

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

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

and for $p - j$ even

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

Here, $s_p = (S^{-1}\tilde{c}_p - S^{-1}\tilde{S}S^{-1}c_p)$, with kernel moment matrices S, $\tilde{S}$, and vectors c_p, $\tilde{c}_p$, for the detailed form of which we refer to Einbeck and Augustin (2005), as well as for the proof, regularity assumptions, and for the asymptotic variance.

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 bias reduction. The augend in the squared bracket in (2) vanishes for $\alpha'(x)/\alpha(x) + f'(x)/f(x) = 0$, and this differential equation is solved for

$$\alpha_{opt}(x) = c\frac{1}{f(x)}, \quad (3)$$

with $c \in \mathbb{R} \setminus \{0\}$. Considering the design density as "selection probability distribution", this gives a similar message to that of HT, where we had optimal weights $\alpha_i = 1/\pi_i$. In practice $f(\cdot)$ is mostly unknown, but it may be substituted by a density estimate, $\hat{f}(\cdot)$.

Design weighted local "robust" smoothing

Formula (3) is exactly the opposite of the recommendation given by Einbeck, André and Singer (2004), who proposed the setting

$$\alpha(\cdot) = \hat{f}(\cdot) \quad (4)$$

in order to robustify against outliers in the design space.

To illustrate the importance of such techniques, we consider the data set analyzed by Conceição et al. (2001) and Singer et al. (2002) to evaluate the association between mortality of children under five attributed to respiratory causes and the concentration of PM_{10}, SO_2, O_3 and CO in the city of São Paulo, Brazil, from 1994 to 1997 (the data is available in `www.ime.usp.br/~jmsinger`).

Fig. 2: Air pollution in São Paulo (Brazil).

The number of daily respiratory deaths as a function of the SO_2 concentration is depicted in Figure 3. Days with high pollutant concentrations (as compared to the majority of the data) are clearly identified. The effect of such observations is to "pull" the fitted curve downward (black line), suggesting that the effect of the pollutant on children mortality decreases for concentrations beyond 50 $(\mu g/m^3)$, a fact that has no biological plausibility. To better understand the effect, observe the two data points at the lower right side of the picture. Although they do not seem to correspond to vertical outliers, they definitely disturb the local linear fit. A possible reason for this is that high concentrations of the pollutant are not coupled with a large number of deaths, contrary to what is expected; this is probably due to the sparse design for large concentrations. It seems clear that some robust method must be employed to bypass this inconsistency. In the light of this example, Einbeck et al. (2004) proposed to downweight the effect of outliers in the explanatory variables using the weights (4). Application of this method to the given data yields the red line in Figure 3.

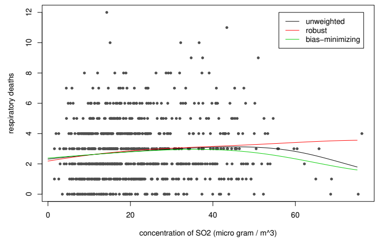

Fig. 3: Respiratory deaths versus SO_2 concentration, local linear fit (black), fit with weights $\alpha \sim \hat{f}$ (red) and $\alpha \sim 1/\hat{f}$ (green).

Tying up the ends

- It is well known that points near the boundary can have a huge influence on the estimate of the regression function. This effect will gain still in power if we apply weights *inversely* proportional to the design density as suggested by our bias-minimizing criterion above - just as Jumbo had a tremendous influence when selected! For instance, in Fig. 3, we see that the green line - i.e. the bias-minimizing fit - is even stronger influenced by the few observations with high SO_2 concentration than the unweighted fit, indicating an even stronger reduction of mortality with rising SO_2 concentration!

- It is worth to look into the rejoinder of Basu's (1971) essay, in which he vehemently denied that the 'unrealistic sampling plan' was responsible for the failure of the HT estimator. Basu defended, in contrary, the circus statistician's sampling plan, as it ensures a *representative* sample, and gave 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.'

- Similarly, in the smoothing context, we have derived a bias-minimizing criterion, which may prove useful for large and well-behaved data sets, but may give disastrous results in the presence of outlying predictors. This is exactly Basu's dilemma: he did not conform himself to the fact that one has to get the selection probabilities right, 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 has to work with it. In the smoothing context, the selection probabilities correspond to the design density, which is almost never designed to meet any optimality criterion, and hence there is always a certain potential that things may go wrong.

- How would such an optimality criterion look like? Applying the bias-minimizing weights (3), one easily verifies that the asymptotic variance (see Einbeck & Augustin, 2005) is minimzed for the design
$$f(x) = c_1 \cdot \sigma^2(x),$$
($c_1 \in \mathbb{R} \setminus \{0\}$). This leads to the "optimal design weights" $\alpha(x) \sim 1/\sigma^2(x)$, which is a weighting scheme well known in the context of parametric regression. However, we do not want to overvalue this result, as the minimized variance term is just a second-order term.

Conclusion

We have demonstrated that there exists a striking analogy between the theories of sampling and smoothing, leading to a similar discrepance between theoretically optimal and practically useful weighting schemes. We believe that this tells us an important lesson about statistical methods in general: weighting is performed in virtually all statistical disciplines, and a usual way of motivating such weights is to look at theoretical, bias-minimizing criteria. These criteria will often suggest to choose weights inversely proportional to some kind of selection probability (density). This however makes the estimator extremely sensitive to extreme observations (corresponding to Jumbo and high SO_2 values, respectively, in our examples).

Hence, we advise users to be careful with bias-minimizing estimators if there are any observations which might be labelled by the terms "extreme", "undesired", "outlying", "weak" or "needy", and the like, and it is likely that this holds far beyond the scope of sampling and smoothing.

References

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

(2) Brewer, K. (2002). *Combined survey sampling inference.* Arnold: London.

(3) Conceição, G.M.S., Miraglia, S.G.E.K., Kishi, H.S., Saldiva, P.H.N. and Singer, J.M. (2001). Air pollution and children mortality: a time series study in São Paulo, Brazil. *Environ Health Perspect* **109**, 347–350.

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

(5) Einbeck, J., and Augustin, T. (2005). On weighted local fitting and its relation to the Horvitz-Thompson estimator. SFB386 Discussion Paper No. 465, University of Munich.

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

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

(8) Singer, J.M., André, C.D.S., Lima, P.L. and Conceição, G.M.S. (2002). Association between atmospheric pollution and mortality in São Paulo, Brazil: Regression models and analysis strategy. In Y. Dodge (Ed), *Statistical Data Analysis based on the L1 norm and related methods*, pp 439–450. Birkhäuser, Berlin.