Smoothing, sampling, and Basu's elephants

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

[1] Department of Math. Sciences, Durham University, Durham DH1 3LE, UK
[2] Institut für Statistik, Ludwigstr. 33, 80539 München, Germany
[3] Departamento de Estatística, Universidade de São Paulo, CP66281, SP, Brazil

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.

Keywords: Weighting; Horvitz-Thompson estimator; local polynomials

1 Introduction

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.
This is a short version of a fable told by Basu (1971), illustrating 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. 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 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 was the total weight of the herd measured at that time (Koop, 1971, in the discussion of Basu's essay).

2 Design-weighted local smoothing

One of the statistical fields where weighting is quite common is that of nonparametric smoothing. Given a sample $(x_1, y_1), \dots, (x_n, y_n)$ 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

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

From the vector $(\hat{\beta}_0(x), \dots, \hat{\beta}_p(x))$ minimizing (1), one easily gets estimators of m and its derivatives, $\hat{m}^{(j)}(x) = j!\hat{\beta}_j(x)$, and one has the following

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*

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

and for $p - j$ even

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

with $s_p = (S^{-1}\tilde{c}_p - S^{-1}\tilde{S}S^{-1}c_p)$, and 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 reduce the bias. Note that 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

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

with $c \in \mathbb{R} \setminus \{0\}$. Considering the design density as "selection probability distribution", this gives a very 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, $\hat{f}(\cdot)$.

3 A surprising analogy

Formula (3) is exactly the opposite of the recommendation given by Einbeck, André and Singer (2004), who proposed the setting $\alpha(\cdot) = \hat{f}(\cdot)$ in order to robustify against outliers in the design space. It is well known that points near the boundary can have a huge influence on the estimate of the regression function (which is even more true for the derivative estimates, see Newell and Einbeck, 2007). This effect will gain dramatically in power if we even 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!

It is at this point worth to take a 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 desastrous results in the presence of outlying predictors. This is exactly the dilemma that Basu was worried about: he did not conform himself to the fact that one has to get the selection probabilites 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 (we are aware that there exist some techniques to adjust the probabilities ex post, e.g., the Keyfitz (1951) technique, which however have drawbacks as they are actually based on deleting available data). 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.

4 Conclusion

The goal of this paper was to show that there exists an 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 (which correspond to Jumbo in Section 1 and the outlying predictors in Section 2). 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" or "needy", and the like, and it is likely that this holds far beyond the scope of sampling and smoothing.

References

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.

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

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.

Newell, J., and Einbeck, J. (2007). A comparative study of nonparametric derivative estimators. *Proceedings of the 22th IWSM*, Barcelona.

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

Keyfitz, N. (1951). Sampling with probabilites proportinal to size: adjustment for changes in the probabilities. *JASA*, **46**, 105–109.

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