The Multihalver

S. Morgenthaler[1]

[1] Ecole polytechnique fédérale de Lausanne, FSB - Institut de mathématiques, 1015 Lausanne, Suisse

Keywords: Jackknife, Inference for Eigendirections, Outlier Detection

1 The Jackknife

The talk will start with a few comments on the jackknife procedure with its advantages and its drawbacks. We will then focus on the multihalver version of the jackknife, which is based on repeated splits of the data into halves. The number of all possible halvings is $\binom{n}{n/2}/2 \sim 2^n/\sqrt{2\pi n}$, which grows quickly with the sample size n. Let $\mathcal{H}$ be the set of halvings to be used in the computation and let T be a real-valued statistic. For each halving $h \in \mathcal{H}$ we compute the difference $T_L - T_R$ of the statistic T, where T_L and T_R are the values of T on the left and right half-samples of that halving, respectively. For each h, we furthermore define a pseudovalue

$$T_h^* = 2T - \left(\frac{T_L + T_R}{2}\right) .$$

The multihalver estimate associated to T is then

$$T_{\mathrm{MH}} = \frac{1}{H} \sum_{h \in \mathcal{H}} T_h^* ,$$

where H is equal to the total number of halvings. The multihalver estimate for the standard deviation of T is

$$\mathrm{Se}_{\mathrm{MH}} = \Big(\frac{1}{H} \sum_{h \in \mathcal{H}} \frac{(T_L - T_R)^2}{4}\Big)^{1/2} .$$

Special care should be taken in the choice of $\mathcal{H}$. If possible each pair of observations should be split between L and R about evenly, that is the four possibilities, both appearing in L, both appearing in R, one in L and the other in R and vice versa should be equally often generated by the halvings in $\mathcal{H}$. We will discuss the use of orthogonal arrays to achieve this and comment on its consequences.

2 Inference for Eigenvectors (joint work with J.W. Tukey)

Two applications of the multihalver will be considered. First, we study confidence cones for the direction of the eigenvector with the larger associated eigenvalue of a two-by-two covariance matrix. This seems to be the simplest truly multivariate inference problem. The performance will – and necessarily has to be – evaluated by simulation in a variety of bivariate distributions, including non-elliptic ones. We need to evaluate a multivariate statistical method on a variety of bivariate distribution shapes if we expect the results to be widely useful. We are not likely to be concerned with sample sizes below 50 (or below 100). Skewness and elongation are the major challenges from one-dimensional causes, and multipole structure is the most obvious two-dimensional consideration. We shall set skewness and multipolarity aside, so far as the present talk goes, partially in view of its possible cure by reexpression, partially because we find it unclear how best to model it, and partially because we have little experience with it in practical examples. Accordingly we shall focus on elongation of our bivariate distributions. Elliptical bivariate distributions differ from circular bivariate ones by only the aspect (part of which) we are to study – stretch in some direction compared to a direction at right angles.

We will show that in order to make progress in the practice of multivariate data analysis, a broader approach than the one based on well-behaved elliptical distributions should be taken. The combined result of applying a variety of small sample adjustments to asymptotic variance estimators, using jackknife estimates, and relying on the robustification leads to good results. How to handle nonelliptical behavior is less clear and is not the main focus of the talk.

3 Outlier Nomination (joint work with L.T. Fernholz and J.W. Tukey)

To identify outliers using the multihalver we will be interested in studying the differences $T_L - T_R$ for all the halvings that we can reasonably obtain. We will propose simple algebraic operations to enrich the basic $\mathcal{H}$. We will then form couples among the observations and proceed to compute outlyingness indicators based on the differences $T_L - T_R$ with the first element of one of the couples always in L and the second element of another couple always in R. An additive analysis of these indicators will then allow us to nominate outliers among the observations. Outlyingness will thus be implicitly defined with respect to the statistic T.

4 References

Fernholz, L. T., Morgenthaler, S. and Tukey, J. W. (2002). An Outlier Detection Method based on the Multihalver. submitted.

Morgenthaler, S. and Tukey, J. W. (1995). Inference for the Direction of the Larger of Two Eigenvectors: The Case of Circular Elongation. In: H. Rieder, editor, *Robust Statistics, Data Analysis, and Computer Intensive Methods: In Honor of Peter J. Huber's 60th birthday*, pp 321–352, Springer Verlag, Heidelberg.