Robust Bootstrap: Influence Function Approach

A. M. Pires[1] and C. Amado[1]

[1] Departamento de Matemática and CMA, Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisboa, Portugal

Keywords: Bootstrap, Influence function, Robust estimation, Outliers.

1 Introduction

The existence of outliers in a sample is an obvious problem which can become worse when the usual bootstrap is applied, because some resamples may have a higher contamination level than the initial sample. Bootstrapping using robust estimators may be a solution to this problem. However, in many instances, this will not be enough because it can lead to several complications, such as: i) the breakdown point for the whole procedure may be small even when based on an estimator with a high breakdown point (Stromberg, 1997; Singh, 1998); ii) mathematical difficulties; iii) very high computation time. In order to solve these problems, we suggest a modification of the bootstrap procedure which consists of forming each bootstrap sample by resampling with different probabilities so that the potentially more harmful observations have smaller probabilities of selection. The aim is to protect the whole procedure against a given number of arbitrary outliers.

2 Robust Bootstrap

As far as we know three authors have addressed this matter previously. Stromberg (1997) studies alternative bootstrap estimates of variability for robust estimators. Singh (1998) suggests a robustification of bootstrap via winsorization for certain L and M estimators and presents a general formula for computing the breakdown point for the p*th* bootstrap quantile of a statistic (a practical difficulty of this method is the winsorization of multivariate samples). Salibian-Barrera and Zamar (2000) introduce a robust bootstrap based on a weighted representation for MM-regression estimates.

Our suggestion for the bootstrap robustification is also to introduce a control mechanism in the resampling plan, consisting of an alteration of the resampling probabilities, by ascribing more importance to some sample values than others and using the influence function to compute those selection probabilities. In general, this procedure leads to resampling less frequently highly influent (in the sense of Hampel's influence function, Hampel et al., 1986) observations while, at the same time, resampling with equal probabilities the observations belonging to the main structure. We assume that the actual distribution of the data belong to a contamination "neighbourhood" of a certain specified "central" parametric model, F_θ. On this context it is necessary to define a robust standardized empirical influence function (with a high breakdown point, in order to avoid masking, and not depending on the observations' scale).

Consider two estimators of θ, a robust one, $\hat{\theta}^r$, and a non-robust one, $\hat{\theta}^{nr}$, which can both be represented by Fisher consistent functionals. We first define the Standardized Influence Function (SIF) of $\hat{\theta}$ (either $\hat{\theta}^r$ or $\hat{\theta}^{nr}$) as

$$SIF(\mathbf{x};\hat{\theta},F_\theta) = \left[IF\left(\mathbf{x};\hat{\theta},F_\theta\right)^T V^{-1}(\hat{\theta},F_\theta) IF\left(\mathbf{x};\hat{\theta},F_\theta\right)\right]^{1/2} \tag{1}$$

where IF denotes the theoretical influence function and

$$V\left(\hat{\theta},F_\theta\right) = E_{F_\theta}\left[IF\left(\mathbf{x};\hat{\theta},F_\theta\right) IF\left(\mathbf{x};\hat{\theta},F_\theta\right)^T\right] \tag{2}$$

is the asymptotic variance of $\hat{\theta}$. Let us assume that, as usual, $SIF(\mathbf{x}; \hat{\theta}^r, F_\theta)$ $(SIF(\mathbf{x}; \hat{\theta}^{nr}, F_\theta))$ depends on F_θ only through a vector of unknown parameters, Ω_1, (Ω_2), and that it has certain invariance properties.

We can then define two robust empirical influence functions by plugging in SIF robust estimates, $\hat{\Omega}_1^r$ or $\hat{\Omega}_2^r$ of the unknown parameters Ω_1 or Ω_2, which will be denoted by $RESIF(\mathbf{x}; \hat{\theta}^r, \hat{\Omega}_1^r)$ or $RESIF(\mathbf{x}; \hat{\theta}^{nr}, \hat{\Omega}_2^r)$. (Two other functions of the same kind could be defined by using non robust estimators of Ω_i, $i = 1, 2$. This does not obviously make sense for $\hat{\theta}^r$. For $\hat{\theta}^{nr}$ it is well known that it would suffer from masking.) From the two defined *RESIF's* only the second is useful to our method. Since most robust estimators have bounded theoretical influence functions it remains impossible to detect dangerous observations (that is, observations that are not harmful when considered alone, but which may cause the collapse of the bootstrap procedure by appearing too often). However, those observations will be easily recognised by a high value of the second *RESIF*. Pison, Rousseeuw, Filzmoser and Croux (2000) in a different context define similar, but non-standardised, empirical influence functions and also recommend the use of the one corresponding to our second. If, for example, we consider multivariate location with multivariate normal distribution as central model then $RESIF(\mathbf{x}; {T_n}^{nr}, \hat{\Omega}_2^r)$, with ${T_n}^{nr} = \bar{\mathbf{x}}$, is simply the robust Mahalanobis distance currently used for (robust) outlier detection in multivariate data sets.

We are now able to present the Influence Function Bootstrap (IFB) procedure. Let $\mathbf{X}_n = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ be an available sample (uni or multivariate), T_n an estimator of a population characteristic, θ, and t_n its sample value. Obtain $RESIF(\mathbf{x}; {T_n}^{nr}, \hat{\Omega}_2^r)$ at each data point: $RESIF_i = RESIF(\mathbf{x}_i; {T_n}^{nr}, \hat{\Omega}_2^r)$, $i = 1, 2, \ldots, n$. Then compute weights, w_i, according to

$$w_i = I_{[0,c]}\left(|RESIF_i|\right) + \psi\left(c, |RESIF_i|\right) \times I_{]c,+\infty]}\left(|RESIF_i|\right), \quad i = 1, 2, \ldots, n$$

where $c > 0$ is a tuning constant and $\psi \geq 0$ is a function verifying $\lim_{t\to\infty} t^2\psi(c,t) = 0$ (for fixed c). Finnally get the resampling probabilities, $p_i = w_i / \sum_{j=1}^n w_j$ $(i = 1, 2, \ldots, n)$. The choice of ψ, the determination of c and some theoretical aspects will be discussed.

3 Applications

The results of Monte Carlo studies comparing the performance of the proposed method, the winsorized bootstrap and the ususal bootstrap, will be presented for the following situations: bootstrap point estimates and confidence intervals for univariate location and for the correlation coefficient, and selection of variables in two-group linear discriminant analysis. Applications to real data sets will also be presented.

References

F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel (1986). *Robust Statistics: The Approach based on Influence Functions.* Wiley & Sons, New York.

G. Pison, P.J. Rousseeuw, P. Filzmoser, and C. Croux (2000). A robust version of principal factor analysis. In: J.G. Bethlehem, and P.G.M. Van der Heijden, editors, *Proceedings in Computational Statistics, Compstat 2000*, pp. 385–390, Physica-Verlag, Heidelberg.

M. Salibian-Barrera and R.H. Zamar (2000). Robust inference and bootstrap. Submitted.

K. Singh (1998). Breakdown theory for bootstrap quantiles. *The Annals of Statistics,* 26, 1719–1732.

A.J. Stromberg (1997). Robust covariance estimates based on resampling. *Journal of Statistical Planning and Inference,* 57, 321–334.

Please fill in this form and mail it together with your abstract.

My abstract fits best to topic number 12 (or to another topic called Resampling)

List of Topics:

1. Algorithms
2. Applications
3. Biostatistics
4. Computing and graphics
5. Data analysis
6. Data mining
7. Economics, finance
8. Efficiency and robustness
9. Functionals and bias
10. Fuzzy statistics
11. Geostatistics
12. Inference for robust methods, model testing
13. Location depth and regression depth
14. Multivariate methods
15. Neural networks
16. Rank-based methods
17. Regression quantiles, trimming
18. Robust covariance
19. Robust designs
20. Robust regression
21. Time series analysis
22. Wavelets
23. Other (please specify)