ROBUST BOOTSTRAP WITH NON RANDOM WEIGHTS BASED ON THE INFLUENCE FUNCTION

Conceição Amado and Ana M. Pires
Department of Mathematics and Applied Mathematics Center
IST, Technical University of Lisbon
Avenida Rovisco Pais - 1049-001, Lisboa, Portugal
camado@math.ist.utl.pt; ana.pires@math.ist.utl.pt

Key Words: bootstrap; influence function; robust estimation; outliers.

ABSTRACT

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; ii) mathematical difficulties; iii) very high computation time. In this paper, we suggest a modification of the bootstrap procedure in order to solve these problems 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. As an illustration, we consider point and interval estimation for the correlation coefficient. We use Monte Carlo methods to compare this method with another robust bootstrap procedure, the winsorized bootstrap, recently suggested by Singh (1).

1. INTRODUCTION

In several statistical problems, it is impossible to derive analytical solutions or it is not possible to establish the assumptions allowing that derivation. In these situations, simulation plays an important role allowing the computation of approximate solutions to those problems. The increasing computation power has been responsible for the widespread use and development of computer-intensive methods. One of these techniques is the bootstrap proposed by Efron (2), where usually the Monte Carlo approximation is adopted, since analytical methods may not be easy to apply. However, when there are outlying cases, the resampling procedure may fail. In this paper we propose a modification of the bootstrap procedure adapted either to samples that have apparently aberrant points or to “good” samples. The paper is organized as follows - in section 2, the general problem of bootstrapping when there are outliers is discussed and the approaches to perform a bootstrap robustification are summarised; in section 3 we present our suggestion; a simulation study is presented in section 4 and the conclusions are drawn in the last section.

2. THE BOOTSTRAP AND OUTLIERS

When a resampling calculation is performed, an important question is: What would be different if one or more observations are removed from the original sample? In the presence of a parametric model there are several diagnostic techniques for detecting and assessing the importance of (usually single) observations. However, when using non-parametric bootstrap what we have is solely the data and nothing more than the data. Therefore, it is very important to look carefully at the simulation results, and verify if substantial differences occur when some observations are removed from the sample. A single case diagnostic, called jackknife-after-bootstrap, was proposed by Efron (3) and has the advantage that the additional effort of simulation can be avoided (see (4)). But the diagnostics may fail (both in parametric or non-parametric calculations) due to the so called masking effect. Even if it does not fail, a question remains: how shall we deal with the possible outliers? The wise attitude is never to discard them before a careful exam in order to find their origin. The application of robust methods easily accomodates this situation since robust estimation is designed to produce estimates that are immune to serious distortion up to a certain number of outliers. However, in what concerns the bootstrap procedure this is not enough because it can lead to the following problems: i) the breakdown point for the whole procedure may be small even when based on an estimator with a high breakdown point (see (5) and (1)) ii) mathematical difficulties, like singularity of matrices; iii) very high computation time, since robust estimates are usually obtained by iterative procedures. It is therefore important to develop bootstrap procedures that prevent some or all of these problems.

As far as we know three authors have addressed this problem previously. Stromberg (5) studies alternative bootstrap estimates of variability for robust estimators. He recommends the so-called SBS robustification of the bootstrap, (that is, bootstrapping as usually but using a 50% breakdown S-estimate of variability instead of the sample variance for the computation of the bootstrap variance estimate) and the so-called JK50 robustification of the jackknife, that is, the deleted 50% jackknife as defined in Shao (6). Note that Stromberg's emphasis is on estimating the variance of an estimator in a robust way, not on a general method for bootstrap robustification. Singh (1) 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 in the application of the **W**insorized **B**ootstrap (WB) is the winsorization of multivariate samples. Salibian-Barrera and Zamar (7) introduce a robust bootstrap based on a weighted representation for MM-regression estimates. They present the asymptotic properties of the method and compute the breakdown point of the quantile estimates derived from it.

3. THE IFB METHOD

Our suggestion for the bootstrap robustification is also to introduce a control mechanism in the resampling plan but different than trimming or winsorizing. The aim is to protect the bootstrap procedure against a given number of arbitrary outliers and at the same time to be computationally less time-consuming than resampling using a robust estimator. The control mechanism consists of an alteration of the resampling probabilities 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, see (8)) 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" model, F_θ.

In the literature there are several definitions of empirical influence function, see for instance Hampel, Ronchetti, Rosseeuw and Stahel ((8) p. 93) and Davison and Hinkley ((4) p. 46). Despite the differences all of them are trying to measure, in some way, the influence that one observation has on the value of an estimate. On the context of this work it was necessary to define and use a robust standardised empirical influence function (with a high breakdown point, in order to avoid masking, and not depending on the observations' scale).

Definition 1: (**R**obust **E**mpirical **S**tandardised **I**nfluence **F**unction (*RESIF*))

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 **S**tandardized **I**nfluence **F**unction (*SIF*) of $\hat{\theta}$ (either $\hat{\theta}^r$ or $\hat{\theta}^{nr}$) as

$$SIF(\mathbf{x};\hat{\theta},F_\theta)=\left[IF\left(\mathbf{x};\hat{\theta},F_\theta\right)^T V^{-1}(\hat{\theta},F_\theta)IF\left(\mathbf{x};\hat{\theta},F_\theta\right)\right]^{1/2} \tag{3.1}$$

where IF denotes the theoretical influence function and

$$V\left(\hat{\theta},F_\theta\right)=E_{F_\theta}\left[IF\left(\mathbf{x};\hat{\theta},F_\theta\right)IF\left(\mathbf{x};\hat{\theta},F_\theta\right)^T\right] \tag{3.2}$$

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).

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 $RESIF(\mathbf{x};\hat{\theta}^{nr},\hat{\Omega}_2^r)$.

Remark 1: 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.

Remark 2: 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 (9) in a different context define similar, but non-standardised, empirical influence functions and also recommend the use of the one corresponding to our second.

Remark 3: It is also important that the robust estimator $\hat{\Omega}_2^r$ has a very high breakdown point, inwhat it will condition the whole procedure.

Remark 4: The *SIF* in equation (3.1) is well defined only if (3.2) is non-singular. When this is not the case it is possible to give an alternative definition. One example is given in Boente, Pires and Rodrigues (10).

Remark 5: It is crucial that the SIF and the $RESIF$ be invariant for certain groups of transformations, depending on the specific model under consideration.

Remark 6: As an example consider multivariate location with multivariate normal distribution as central model, that is, $F_\theta(\mathbf{X}) = F_0\left(\mathbf{\Sigma}^{-1/2}(\mathbf{X}-\theta)\right)$, with $F_0 \sim N_p(\mathbf{0}, \mathbf{I})$, and allowing for actual distribution of the data one in the gross-error neighbourhood

$$H_\varepsilon = \{H = (1-\varepsilon)F_\theta + \varepsilon H^* : H^* \text{ is any distribution in } R^p\}.$$

If $T_n^{nr} = \bar{\mathbf{X}}$ (the optimal estimator under the central model) then $\Omega_2 = (\theta, \mathbf{\Sigma})$ and let $\hat{\Omega}_2^r = (\hat{\theta}^r, \hat{\mathbf{\Sigma}}^r)$ denote robust estimators of Ω_2. It is easy to verify that $RESIF(\mathbf{x}; T_n{}^{nr}, \hat{\Omega}_2^r) = (\mathbf{x}-\hat{\theta}^r)^T\left(\hat{\mathbf{\Sigma}}^r\right)^{-1}(\mathbf{x}-\hat{\theta}^r)$, which is simply the robust Mahalanobis distance currently used for (robust) outlier detection in multivariate data sets, and which is invariant under the group of affine transformations.

Now, we are able to present the **I**nfluence **F**unction **B**ootstrap (IFB) procedure. The general method is described on algorithm 1. The second algorithm is just the general description of the bootstrap. Let $\mathbf{X}_n = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ be an available sample (uni or multivariate), T_n an estimator of a population characteristic, θ, and t_n its sample value.

Algorithm 1 (IFB)

(computation of the resampling probabilities $\mathbf{P} = (p_1, p_2, \ldots, p_n)$)

Step 1) Obtain the $RESIF(\mathbf{x}; {T_n}^{nr}, \hat{\Omega}_2^r)$

Step 2) Compute the estimates of the $RESIF(\mathbf{x}; {T_n}^{nr}, \hat{\Omega}_2^r)$ 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

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

where I_A is the indicator function of the set A, $c > 0$ is a tuning constant and ψ is a non-negative function verifying the following two conditions:

Condition 1) $\lim\limits_{t \to +\infty} t^2 \psi(c, t) = 0$ (for fixed c)

Condition 2) $\frac{\partial}{\partial t} \psi \left(c, t \right)_{|t=c} = 0$

Step 3) Get $\mathbf{P} = (p_1, p_2, \ldots, p_n)$ where $p_i = w_i / \sum_{j=1}^{n} w_j$.

Algorithm 2

(bootstrap with resampling probabilities $\mathbf{P}$)

Step 1) Let $G_{n\mathbf{P}}$ be a distribution function assigning probabilities $p_1, p_2, \ldots, p_n$ to $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$.

Step 2) Resample with replacement according to $G_{n\mathbf{P}}$. Let $\mathbf{X}_n^* = (\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_n^*)$ be the current bootstrap resample determined and compute the bootstrap estimate, $t_n^* = T_n(\mathbf{X}_n^*)$.

Step 3) Repeat step 2 independently B times, to get $t_n^{*_1}, t_n^{*_2}, ..., t_n^{*_B}$ (bootstrap replicates which determine the bootstrap empirical distribution of the estimator T_n).

The following proposition justifies the need of condition 1 on Algorithm 1.

Proposition 1: *In order to protect the bootstrap distribution from the harmful effect of outliers the function $\psi(c,t)$ must be $o(t^{-2})$ as $t \to +\infty$.*

Proof:

Let be $\mathbf{X}_n = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ be a sample of size n.
Suppose that $\#\{x_i : x_i \text{ is an outlier}\} = k < \frac{n}{2}$ and for simplicity's sake all at the same point, ξ_{out}. If $P(\xi_{out} \text{ being resampled}) = p$ with $p \leq \frac{1}{n} < \frac{1}{k}$ then the random variable X*: number of outliers in the resample* will follow a $Bin(n, kp)$ distribution.

We then choose a high percentile, b_α, with $\alpha \approx 1$, of this distribution such that only a few resamples will have more than b_α outliers (that is, only the extreme percentiles of the bootstrap distribution will be affected). b_α will be of order $nkp + z_\alpha\sqrt{nkp}$ (assuming kp is much smaller than 1) where $z_\alpha = \Phi^{-1}(\alpha)$. Then the k outliers will affect the smallest α percentiles of the bootstrap distribution by at most $b_\alpha t$ where $t = IF(\xi_{out}; T_n, F)$.

Suppose that p is of the order of $|t|^{-1}$, then $b_\alpha t$ is still unbounded.
Now, suppose that p is of the order of $|t|^{-2}$ then $b_\alpha t$ is bounded but $\lim_{t \to +\infty} b_\alpha t \neq 0$, which means that the outliers may introduce a severe bias. Therefore p must decrease faster than $|t|^{-2}$. $\diamond$

The second condition for ψ in algorithm 1 is just a smoothness condition introduced in order to preserve the efficiency of the procedure with clean data or with moderate outliers. It can also be viewed as a smooth rejection of outliers in opposition to hard rejection.

A flexible family of functions from where the ψ function can be chosen is the kernel of the p.d.f. of the t-distribution and its limiting form, the normal distribution:

$$\Psi(x; c, d, \gamma) = \begin{cases} \left[1 + \frac{(x-c)^2}{\gamma d^2}\right]^{-\frac{\gamma+1}{2}} & 0 < \gamma < \infty \\ exp[-\frac{(x-c)^2}{2d^2}] & \gamma = \infty \end{cases}$$

where c is the location parameter, d is the scale parameter and γ is a shape parameter. Condition 1 is satisfied if $\gamma > 1$ and condition 2 if the location parameter is the tuning constant. Using this family introduces two more parameters to adjust besides the tuning constant (since the scale parameter does not affect the decreasing rate, in the example described in the next section, we have reduced the number of parameters to adjust by setting $d = c$).

Other kinds of empirical influence functions could be used, for instance, those provided by the jackknife (see for instance (4) p. 50) and in this case step 1 of Algorithm 1 is not necessary. However, this kind of empirical influence function may not reflect in a good way the outlying cases, as it may suffer from masking. Furthermore, it will be computationally more expensive because IFB requires the weights before the final bootstrap.

With this resampling plan, we can guarantee that a large amount of the bootstrap replicates will be free of the harmful influence of the outliers. However, there is still the possibility that the most extreme values are affected. This is not important if we want to use only percentiles of the distribution, for instance for interval estimation, but it may affect the empirical mean or variance. Therefore, these quantities must be replaced by robust measures of the same type, like the median or the trimmed mean and the scaled interquantile range or a trimmed variance.

The adjustment of the tuning constant can be solved by a prior simulation study. That is, simulate the distribution of *SIF* at a central parametric model and choose for c a very high percentile of this distribution. As usual with tuning constants in robust estimation there is a trade-off between efficiency and robustness. Higher c will lead to a more efficient procedure (again at a certain central model) but will also be less robust. On the other hand, a lower value of c will ensure a more robust but less efficient procedure. Note that by choosing a very high percentil the IFB procedure will be "almost" equivalent to the usual bootstrap at the central model. The following considerations may be used for choosing the very high percentil mentioned. Let $p_c = P(SIF > c)$, then the probability of having to downweight at least one observation in an "clean" sample of size n is $p_{c,n} = 1 - (1 - p_c)^n$. If Y is the number of bootstrap samples with at least one downweighted observation then $Y \sim Binomial(B, p_{c,n})$. Finnaly, we may fix a risk, q, of affecting a given proportion, a, of the bootstrap samples, that is, $P(Y \geq aB) = q$. If for instance $a = 0.10$, $q = 0.05$ and $n = 100$ straightforward calculations lead to $p_{c,n} = 0.084785$ and $p_c = 0.0009$. Note that this calculations do not consider the size of the weights and, as the ones choosen decrease very smoothly, the actual effect, though present, may be very small.

4. SIMULATION STUDY

As an example we applied the procedure described in the previous section to obtain bootstrap confidence intervals for the correlation coefficient, ρ. We chose this example because it appears very often in the bootstrap literature and it is known that the sample product-moment correlation coefficient, r, is very sensitive to outliers.

4.1. DESCRIPTION

4.1.1. DESIGN

A simulation study was conducted to assess the performance of the WB and IFB. Two other methods were used, as control, the **U**sual **B**ootstrap (UB), and Fisher's $tanh^{-1}(r)$. As it is also important to compare these methods with the outlier rejection method, we used the $RESIF(\mathbf{x}; r, \hat{\Omega}_2^r)$ to identify outliers, then we deleted those outliers and bootstraped the remaining data. We denote this method by **R**ejection **B**ootstrap (RB). For the resampling methods, the 0.05 and 0.95 percentile bootstrap limits were obtained.

The "control" distribution was the bivariate Normal, $N_2(\mathbf{0}, \mathbf{\Lambda})$, (NOR); the other distributions were: Random Contaminated Normal, (RCN), $0.75N_2(\mathbf{0}, \mathbf{\Lambda}) + 0.25N_2(\mathbf{0}, 9\mathbf{\Lambda})$; and Deterministic Contaminated Normal (DCN), where 10% of a $N_2(\mathbf{0}, \mathbf{\Lambda})$ sample was replaced by the point $(-5, 5)$. $\mathbf{\Lambda}$ is a bivariate covariance matrix with off-diagonal element ρ and unit diagonal elements. We used samples of size $n = 100$ and $\rho = 0.5$. Two estimators were considered: i) non robust - Pearson correlation coefficient, r; ii) robust - minimum covariance determinant, MCD, (see (11)) in its weighted version. The MCD objective is to look for h observations (out of n) whose classical covariance matrix has the lowest determinant. The choice $h = [(n + p + 1)/2] \approx 0.5n$ leads to the highest breakdown point, which is about 50% for $p = 2$ (see (12)). That is, to cause a bad perturbation on the MCD estimate it is necessary to change about $0.5n$ observations. Usually, the weighted version, which maintains the breakdown point and yields better efficiency, is used. In this study we use the weighted versions of MCD50 ($h \approx 0.5n$) and MCD25 ($h \approx 0.75n$, for which the breakdown point is about 25%). According to remark 2 on page 5 in the case of the MCD estimators and our resampling plan (IFB) we used the weights provided by $RESIF(\mathbf{x}; r, \hat{\Omega}_2^r)$ and not by the $RESIF(\mathbf{x}; MCD, \hat{\Omega}_1^r)$. When bootstrapping the MCD and because of the high computation time required only 100 simulations were performed. For the other situation (r-estimator) the simulation size was 1000. In both cases, for resampling methods, 1000 bootstrap replicates were used. We have used the non-parametric bootstrap despite having set the data distributions.

All computations were done on a Pentium III PC at 550 Mhz using S-plus 4.5.

4.1.2. THE IFB WEIGHTS

The influence function of $T_n = r$ for any bivariate distribution, F, with finite second moments at $\mathbf{x} = (x_1, x_2)^T$ is

$$IF(x_1, x_2; r, F) = -0.5\rho({y_1}^2 + {y_2}^2) + y_1 y_2, \quad (|\rho| \neq 1)$$

where y_1 and y_2 are standardised forms of x_1 and x_2 (see (13)). Therefore the standardised influence function is

$$SIF(x_1, x_2; r, F) = \frac{|IF\,(x_1, x_2; r)\,|}{(1 - \rho^2)} \tag{4.1}$$

with the asymptotic variance computed under the normal distribution. It shall be noted that this does not mean we are assuming the data is normally distributed, it is only a consequence of the optimality of r under the normal distribution, therefore influent observations must refer to this distribution.

The robust empirical standardised influence function ($RESIF$) is obtained when the unknow parameters in (4.1) are replaced by the corresponding MCD50 estimates. Finally the weights are given by

$$w_i = I_{[0,c]}\left(|RESIF_i|\right) + \left[1 + \frac{(|RESIF_i| - c)^2}{\gamma c^2}\right]^{-(\gamma+1)/2} \times I_{]c,+\infty[}\left(|RESIF_i|\right),$$

$i = 1, 2, \ldots, n$, with three possible values for γ, $\gamma = 2, 10, \infty$.

The following proposition states two important invariance properties of the SIF defined in (4.1).

Proposition 2: *Let* $\mathbf{X} = (X_1, X_2)^T$ *denote a random vector with* $E(\mathbf{X}) = \boldsymbol{\mu}$, $Var(\mathbf{X}) = \boldsymbol{\Sigma}$, $\rho = corr(X_1, X_2)$ *and distribution function* F. *Then*

(a) The distribution of $SIF(\mathbf{X}; r, F)$ *does not depend on* ρ.

(b) $SIF(\mathbf{X}; r, F)$ *is invariant under affine transformations of* $\mathbf{X}$.

The proof is given in the Appendix.

This proposition entails that the $RESIF$ and the weights are also affinely invariant, provided the robust estimators used are affinely equivariant.

Three possible values for c, ($c = 5, 7, 10$) were chosen. We verify by numerical integration that these are, respectively, the approximate 0.9977, 0.99977 and 0.99999 percentiles of the SIF under the bivariate normal distribution (which is simply the absolute value of the product of two independent standard normal variates, see the proof of Proposition 2 in the Appendix).

4.1.3. THE BIVARIATE WINSORIZING FOR WRB

The steps used to perform the winsorizing of a $\beta - fraction$ ($\beta = 0.1$) of a bivariate sample were:

(i) Compute robust estimates of the mean vector and covariance matrix;

(ii) Order the sample based on a robust version of the Mahalanobis distance: this is obtained by plugging-in robust estimates;

(iii) Let $[n\beta]$ be the largest integer less than or equal to $n\beta$. The $[n\beta]$ points with largest robust Mahalanobis distance will be moved until their distances are equal to that for the most distant unadjusted point but preserving the orientation information of these points.

The choice of $\beta = 0.1$ is based on the worst case of our contamination distribution, which is DCN with 10%.

4.2. MAIN SIMULATION RESULTS

Table 1 gives estimates of noncoverage probability and descriptive measures (mean and standard deviation) of the interval length, for the three fixed tuning constants and the three values of γ, obtained for IFB with r. We can see from these results that, as expected, the choice of γ is not as important as the choice of c, however, because of a larger interval length for $\gamma = 2$ (at DCN) and a slightly better coverage probability (at NOR) we recommend $\gamma = 10$. The trade-off between efficiency and robustness related to the tuning constant mentioned at the end of section 3 is not easily observed in these results because we would have to compare, either the mean interval length for a fixed attained noncoverage, or the noncoverage for a fixed attained interval length. However, it can be seen that: at the NOR distribution there is a tendency to achieve a better coverage probability for higher c (ignoring the interval length); and at the DCN we can observe an increase in the mean interval length as c increases (ignoring the differences in noncoverage). The reason why these differences are not very important may be related to the fact that all the constants used correspond to very high percentiles.

Table 1: Results of the simulation for different constants with the IFB method.

Sampling distribution	Weight function	(target: 0.100) noncoverage			Interval length Mean			Interval length Standard deviation		
	γ	c=5	c=7	c=10	c=5	c=7	c=10	c=5	c=7	c=10
NOR	∞	0.107	0.119	0.120	0.2433	0.2429	0.2440	0.033	0.033	0.034
	10	0.133	0.130	0.110	0.2427	0.2423	0.2430	0.034	0.034	0.033
	2	0.124	0.109	0.113	0.2434	0.2454	0.2426	0.032	0.034	0.034
RCN	∞	0.222	0.172	0.130	0.2675	0.2893	0.3147	0.049	0.053	0.058
	10	0.214	0.147	0.125	0.2700	0.2957	0.3174	0.048	0.053	0.058
	2	0.194	0.164	0.113	0.2761	0.2985	0.3187	0.047	0.054	0.060
DCN	∞	0.130	0.134	0.130	0.2454	0.2427	0.2501	0.035	0.035	0.047
	10	0.128	0.140	0.116	0.2425	0.2443	0.2889	0.034	0.039	0.084
	2	0.120	0.088	0.089	0.2747	0.3653	0.4925	0.053	0.079	0.087

After this analysis, and taking into account all the aspects, we decided that the best compromise is attained by choosing $c = 10$ and $\gamma = 10$ and we used the results for this pair of parameters for the rest of the simulation where IFB is compared with the other bootstrap methods (UB, WB, and RB).

Table 2 contains the estimates of noncoverage probability and descriptive measures (mean and standard deviation) of the interval length for each method and distribution, when the r estimator was bootstrapped.

Table 2: Results of the simulation with Pearson correlation coefficient,r.

Sampling distribution	Method	(target: 0.100) noncoverage	Interval length Mean	Interval length Standard deviation
NOR	Fisher	0.104	0.2476	0.024
	UB	0.106	0.2433	0.035
	IFB(c=10,$\gamma = 10$)	0.110	0.2430	0.033
	WB	0.166	0.2245	0.028
	RB(c=10)	0.122	0.2422	0.033
RCN	Fisher	0.267	0.2456	0.036
	UB	0.124	0.3498	0.079
	IFB(c=10,$\gamma = 10$)	0.125	0.3174	0.058
	WB	0.165	0.2706	0.045
	RB(c=10)	0.256	0.2547	0.049
DCN	Fisher	1.000	0.2218	0.019
	UB	1.000	0.3669	0.034
	IFB(c=10,$\gamma = 10$)	0.116	0.2889	0.084
	WB	0.926	0.3592	0.041
	RB(c=10)	0.125	0.2563	0.037

Fisher's $tanh^{-1}$ method, used as control, is optimum when there is no contamination but it fails for the other distributions. We can see that UB and WB (even though this last plan has been suggested for robust estimators) have a reasonable performance under the RCN distribution. This may be due to the kind of distribution, which preserves the correlation coefficient, and gives no preferential treatment to any direction, in the variable space, for locating the simulated outliers. But for DCN, UB and WB fail. A remarkable feature of IFB is the relative constancy of its noncoverage probabilities across the distributions considered.

Table 3: Results of the simulation with the MCD estimators.

		(target: 0.10)		Interval length			
				MCD50		MCD25	
Sampling		noncoverage			Standard		Standard
distribution	Method	MCD50	MCD25	Mean	deviation	Mean	deviation
NOR	UB	0.02	0.07	0.5669	0.127	0.3759	0.075
	IFB	0.03	0.07	0.5869	0.139	0.3886	0.071
	WB	0.01	0.05	0.5670	0.133	0.3783	0.084
RCN	UB	0.06	0.09	0.5304	0.124	0.3977	0.085
	IFB	0.05	0.10	0.5409	0.136	0.3906	0.080
	WB	0.02	0.03	0.5844	0.141	0.4314	0.096
DCN	UB	0.03	0.08	0.4739	0.096	0.3346	0.076
	IFB	0.03	0.15	0.6171	0.159	0.3891	0.093
	WB	0.02	0.07	0.9287	0.311	0.6524	0.223

Table 3 contains the same type of results as Table 2 but for the MCD estimators. For MCD50 all methods have a similar performance: all noncoverage probabilities are lower than the target and all the mean interval lengths are much higher than those in Table 2, even for UB, which can be explained by the low efficiency of the MCD50 estimator. For MCD25 an increase of efficiency is observed (that is smaller interval length) for all methods. On the other hand the IFB approach has a reasonable behaviour across the three distributions and at the contaminated distributions is better than WB but it is not better than UB. Note that, because in Table 3 the number of simulations is 100, the uncertainty about the noncoverage probabilities estimates is larger than for those given in Table 2. As an example an approximate 95% confidence interval for this probability, in the case 0.15 (MCD25 with DCN and IFB), is $(0.08; 0.22)$.

There were no mathematical or statistical problems when the MCD estimator was bootstrapped because the sample size was large enough and the MCD estimators have a high breakdown point. The only problem was the high computation time required by this procedure (about 20 times more than bootstrapping with a non robust estimator).

Overall, IFB with the non robust estimator, emerges as the best compromise between efficiency and robustness. (If we would still want to bootstrap these robust estimators and at the distributions considered then UB leads to smaller confidence intervals than IFB or WB).

4.3. INCREASING SAMPLE SIZE

We have also computed bootstrap confidence intervals for the correlation coefficient, ρ using $IFB(c = 10, \gamma = 10)$ and r under the NOR distribution with sample sizes $n = 100, 1000, 5000, 10000$ and 20000. To control we have used Fisher's $tanh^{-1}(r)$. Table 4 shows the results of 1000 simulations for each sample size. It seems that there is no serious undercoverage for IFB with uncontaminated data as n increases, and that the estimated mean interval length, for both methods, is similar and approximately proportional to $1/\sqrt{n}$. There seems to be only a second order effect on the standard deviation of the lengths with IFB (which is approximately 1.5 to 5 times larger than that of Fisher's method, increasing with n).

Table 4: Results of simulations with increasing sample size for uncontaminated data.

			Interval length			
	(target: 0.100)		IFB		Fisher	
sample	noncoverage			Standard		Standard
size	IFB	Fisher	Mean	deviation	Mean	deviation
$n = 100$	0.110	0.104	0.2430	0.0328	0.2476	0.0243
$n = 1000$	0.110	0.118	0.0777	0.0042	0.0781	0.0025
$n = 5000$	0.105	0.106	0.0347	0.0011	0.0348	0.0005
$n = 10000$	0.091	0.089	0.0246	0.0008	0.0247	0.0002
$n = 20000$	0.098	0.097	0.0174	0.0005	0.0174	0.0001

4.4. EMPIRICAL BREAKDOWN POINT STUDY

The finite sample breakdown point is the largest proportion of the n observations which can be changed arbitrarily before the estimator becomes meaningless.

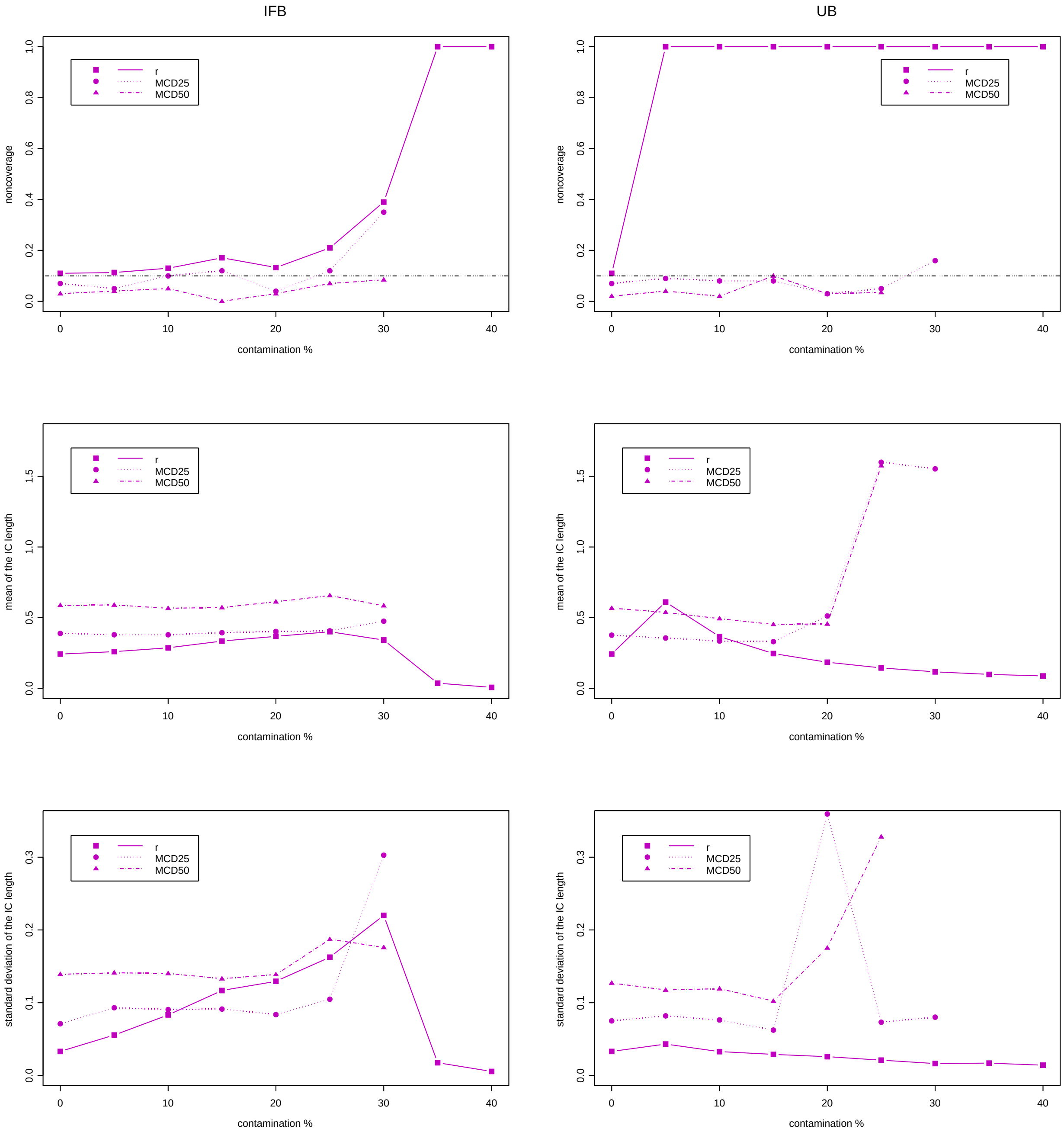

Figure 1: Characteristics of IFB and UB confidence intervals for ρ (noncoverage, mean and standard deviation) against level of contamination.

The process of calculating the breakdown point of IFB in general is nontrivial and is beyond the scope of this paper. In this section we will only estimate it by simulation for the correlation coefficient. 90% confidence limits were obtained for the IFB and UB resampling plans with the three estimators previously considered (r, MCD50, MCD25). We have generated 1000 (for r) and 100 (for MCD50 and MCD25) samples of size 100 from the DCN distribution with the contamination at the point $(-5, 5)$ varying from 0% to 40% (by steps of 5%). Figure 1 shows the simulation results. It shall be noted that it was not possible to obtain all the results for the MCD estimators under the higher contamination levels due to singularity problems.

We can see that the empirical breakdown point for IFB and r is between 30% and 35%, whereas the empirical breakdown point for UB and MCD50 (or MCD25) is between 20% and 25% (since the mean interval length approaches 2, rendering the interval useless). It is again evident from these results that, up to 20% of contamination, the best method is IFB and r.

5. EXAMPLE

We have applied $IFB(c = 10, \gamma = 10)$ to compute a confidence interval for the correlation coeficient between the LSAT and GPA variables of the Law school data (see (14), pag.10). After some diagnostics Efron (3) suggested that, at least, one point (in Efron'paper denoted by A) could be considered "strange".

Two other obvious confidence intervals are computed: i) Fisher's $tanh^{-1}(r)$ without point A and ii) UB method and r without point A.

Table 5 shows the nine 90% confidence intervals for $\rho(LSAT, GPA)$. (Note: it was not possible to bootstrap MCD50, either for UB or IFB, because for a large number of resamples the estimated covariance matrix was singular). All the bootstrap intervals are based on B=1000 nonparametric bootstrap replications. The smallest interval length is obtained with IFB and r. This interval is somehow different from the second and the fourth which shows that there are other points besides A affecting r. Assuming that the data comes from a bivariate normal distribution with some amount of contamination, and in face of the simulation results in subsection 4.2, we claim that this is the best interval. Table 5 also gives point estimates associated with each method. For the bootstrap methods we give two point estimates: the usual one, which is the mean of the bootstrap replicates, and the 5% trimmed mean, as suggested at the penultimate paragraph of section 3.

Table 5: Approximate 90% confidence limits and point estimates for the correlation coefficient ρ (Law school data).

Method	Left limit	Right limit	Point estimation mean	trimmed-5%
Fisher	0.5089	0.9071	$r = 0.7764$	
Fisher without point A	0.7353	0.9589	$r_{-A} = 0.8929$	
UB and r	0.5217	0.9470	0.7697	0.7776
UB and r without point A	0.7787	0.9683	0.8861	0.8902
UB and MCD25	0.3017	0.9982	0.8573	0.8909
IFB and r	0.8679	0.9845	0.9373	0.9405
IFB and MCD25	0.4983	0.9991	0.9163	0.9457
Bootstrap $BC_a(a = 0)$	0.4880*	0.9000*	—	—
Bootstrap-t	−0.0600*	0.9400*	—	—

* From Shao and Tu (15) pag. 169

6. CONCLUSIONS

The idea of robust bootstrap is at the opposite end of traditional and diagnostics approaches, it uses a single bootstrap, which is safe against arbitrary outliers. Sometimes, this protection translates into some loss of statistical efficiency in situations where the samples are outlier free. But this loss can be small as the simulation results show. The method suggested (IFB) is less “automatic” than the winsorized bootstrap of Singh, since it requires analytical computation of the influence function of the estimator and the choice of a tuning constant. However, for many usual estimators influence functions are provided in the literature and the determination of an appropriate tuning constant by simulation or by numerical methods is less time consuming than the bootstrap itself. On the other hand, Singh’s approach is likely to be very difficult to use with multivariate data.

The most interesting aspect of the resampling plan suggested is the ability to compute robust point and interval estimates with a classical estimator, using a robust estimator only in one step and avoiding the computation of bootstrapping a robust estimator.

APPENDIX

Proof of Proposition 2:

Let $\mathbf{Y} = (Y_1, Y_2)^T$ be the standardised form of $\mathbf{X}$. Then $E[\mathbf{Y}] = \mathbf{0}$ and $Var(\mathbf{Y}) = \mathbf{\Lambda}$, with

$$\mathbf{\Lambda} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

We can write $\mathbf{Y} = \mathbf{\Lambda}^{1/2}\mathbf{Z}$, where $\mathbf{Z} = (Z_1, Z_2)^T$ is such that $E[\mathbf{Z}] = \mathbf{0}$ and $Var(\mathbf{Y}) = \mathbf{I}_2$, and

$$\mathbf{\Lambda}^{1/2} = \begin{bmatrix} \sqrt{1+\rho} + \sqrt{1-\rho} & \sqrt{1+\rho} - \sqrt{1-\rho} \\ \sqrt{1+\rho} - \sqrt{1-\rho} & \sqrt{1+\rho} + \sqrt{1-\rho} \end{bmatrix} \times \frac{1}{2}.$$

It is easy to verify that $SIF(\mathbf{X}; r, F) = |\mathbf{Y}^T\mathbf{A}\mathbf{Y}|$, with

$$\mathbf{A} = \begin{bmatrix} -\rho & 1 \\ 1 & -\rho \end{bmatrix} \times \frac{0.5}{(1-\rho^2)}.$$

Hence $SIF(\mathbf{X}; r, F) = |\mathbf{Z}^T\mathbf{\Lambda}^{1/2}\mathbf{A}\mathbf{\Lambda}^{1/2}\mathbf{Z}|$. But

$$\mathbf{\Lambda}^{1/2}\mathbf{A}\mathbf{\Lambda}^{1/2} = \begin{bmatrix} 0 & 0.5 \\ 0.5 & 0 \end{bmatrix}.$$

Therefore $SIF(\mathbf{X}; r, F) = |Z_1 Z_2|$, which completes the proof of part (a) because the distribution of $\mathbf{Z}$ does not depend on ρ.

Let $F_{C,b}$ be the distribution of $\mathbf{X}_{C,b} = \mathbf{C}\mathbf{X} + \mathbf{b}$, where $\mathbf{b} \in R^2$ and $\mathbf{C}$ is a non-singular (2×2) matrix. We want to prove that

$$SIF(\mathbf{x}; r, F) = SIF(\mathbf{x}_{C,b}; r, F_{C,b})$$

for every $\mathbf{x} \in R^2$ and $\mathbf{x}_{C,b} = \mathbf{C}\mathbf{x} + \mathbf{b}$.

From the proof of (a) it is necessary only to show that $\mathbf{Z} = \mathbf{Z_{C,b}}$ with $\mathbf{Z} = \mathbf{\Sigma}^{1/2}(\mathbf{x} - \boldsymbol{\mu})$ and $\mathbf{Z_{C,b}} = \mathbf{\Sigma}_{C,b}^{1/2}(\mathbf{x}_{C,b} - \boldsymbol{\mu}_{C,b})$ where $\mathbf{\Sigma}_{C,b} = \mathbf{C}\mathbf{\Sigma}\mathbf{C}^T$ and $\boldsymbol{\mu}_{C,b} = \mathbf{C}\boldsymbol{\mu} + \mathbf{b}$. Choosing $\mathbf{\Sigma}_{C,b}^{1/2}$ as $\mathbf{C}\mathbf{\Sigma}^{1/2}$ the result follows easily.

BIBLIOGRAPHY

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

(2) Efron, B. Bootstrap methods: another look at the jackknife. The Annals of Statistics, **1979**, *7*, 1-26.

(3) Efron, B. Jackknife-after-bootstrap standard errors and influence functions. Journal of the Royal Statistical Society, **1992**, *54*, 83-127.

(4) Davison, A.C.; Hinkley, D.V. *Bootstrap Methods and their Application.* Cambridge University Press, New York, **1997**.

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

(6) Shao, J. The efficiency and consistency of approximations to the jackknife variance estimators. Journal of American Statistics Association, **1989**, *84*, 114-119

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

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

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

(10) Boente, G.; Pires, A.M.; Rodrigues, I. Influence functions and outlier detection under the common principal components model. Preprint, **2001**, Departamento de Matematica e Instituto de Cálculo. Facultad de Ciencias Exactas y Naturales, Univ. Buenos Aires. Submitted.

(11) Rousseeuw, P.J. Least median of squares regression. Journal of the American Statistical Association, **1984**, *79*, 871-880.

(12) Lopuhaä, H.P.; Rousseeuw, P.J. Breakdown properties of affine equivariant estimators of multivariate location and covariance matrices. The Annals of Statistics, **1991**, 19, 229-248.

(13) Devlin, S.J.; Gnanadesikan, R.; Kettenring, J.R. Robust estimation and outlier detec-

tion with correlation coefficients. Biometrika, **1975**, *62*, 531-545.

(14) Efron, B. *The Jackknife the Bootstrap and Other Resampling Plans.* CBMS 38, SIAM-NSF. **1982**.

(15) Shao, J.; Tu, D. *The Jackknife and Bootstrap.* Springer-Verlag, New York, **1995**.