

# Globally Robust Inference

Jorge Adrover

Universidad Nacional de Córdoba and CIEM,  
Argentina

José R. Berrendero

Universidad Autónoma de Madrid,  
Spain

Matias Salibian-Barrera

Carleton University,  
Canada

Ruben H. Zamar

University of British Columbia,  
Canada

## Abstract

*The robust approach to data analysis uses models that do not completely specify the distribution of the data, but rather assume that this distribution belongs to a certain neighborhood of a parametric model. Consequently, robust inference should be valid under all the distributions in these neighborhoods. Regarding robust inference, there are two important sources of uncertainty: (i) sampling variability and (ii) bias caused by outlier and other contamination of the data. The estimates of the sampling variability provided by standard asymptotic theory generally require assumptions of symmetric error distribution or alternatively known scale. None of these assumptions are met in most practical problems where robust methods are needed. One alternative approach for estimating the sampling variability is to bootstrap a robust estimate. However, the classical bootstrap has two shortcomings in robust applications. First, it is computationally very expensive (in some cases unfeasible). Second, the bootstrap quantiles are not robust. An alternative bootstrap procedure overcoming these problems is presented. The bias uncertainty is usually ignored even by robust inference procedures. The consequence of ignoring the bias can result in true probability coverage for confidence intervals much lower than the nominal ones. Correspondingly, the true significance levels of tests may be much higher than the nominal ones. We will show how the bias uncertainty can be dealt with by using maximum bias curves, obtaining confidence interval and test valid for the entire neighborhood. Applications of these ideas to location and regression models will be given.*

# 1 Introduction

Most results in the robustness literature concern robust point estimates while inference methods have not received the same amount of attention.

The uncertainty of point estimates stems from at least two sources: *sampling variability* (or variance) and *sample quality* (or bias). Sampling variability has constituted the main focus of statistical theory and practice in the past century. The bias caused by poor/uneven data quality, data contamination, gross errors, missing values, etc. has received much less attention.

In our opinion the unbalanced research effort devoted to these two topics does not reflect their actual importance. Since standard errors of point estimates are usually of order  $O(1/\sqrt{n})$  while biases are of order  $O(1)$ , the uncertainty arising from *data quality* (bias) typically dominates that arising from *data quantity* (sampling variability) for moderate and large data sets.

The great attention statisticians paid to the problem of sampling variability is justified by its obvious importance (specially in the case of small samples) and perhaps by the relative ease of this problem: sampling variability can be easily modeled and measured. On the other hand, sample quality and bias are more difficult to model and measure.

Suppose that  $\{x_1, \dots, x_n\}$  are independent and identically distributed random variables with common distribution  $F$  in the family

$$\mathcal{V}(F_0, \epsilon) = \{F : F = (1 - \epsilon)F_0 + \epsilon H, H \text{ arbitrary}\}, \quad (1)$$

where  $0 < \epsilon < 0.5$ ,  $F_0(y) = \Phi((y - \mu_0)/\sigma_0)$ , and  $\Phi = N(0, 1)$ . The robustness model (1), called *Tukey's contamination neighborhood*, is a simple and flexible device to model datasets of uneven quality. According to this model  $(1 - \epsilon)100\%$  of the data follows a normal location-scale model and  $\epsilon 100\%$  of the data comes from an arbitrary, unspecified source. We can choose  $\epsilon$  and  $H$  to represent situations of asymmetry/heavy tails of the error distribution, isolated outliers and cluster of outliers. We can interpret  $(1 - \epsilon)100\%$  as the minimum possible percentage of good quality measurements in our dataset. In summary,  $(1 - \epsilon)$  is a *parameter* measuring the *quality* of the data.

We will show here how the robustness theory developed in the second half of past century can be used to address, at least in part, the problem of performing *globally*

robust inference.

Two aspects of the inference procedure need to be addressed: the asymptotic bias of the point estimates, and the estimation of their distribution for an *arbitrary* distribution in the contamination neighborhood (1).

To fix ideas, let us consider location M-estimates  $T(F_n)$  that satisfy the equation

$$\sum_{i=1}^n \psi((x_i - T(F_n)) / S(F_n)) = 0, \quad (2)$$

where  $\psi$  is an appropriate (monotone) score function and  $S(F_n)$  is a robust scale estimate that converges to  $S(F)$ . Under very mild regularity conditions (see Huber, 1981)  $T(F_n)$  converges to the value  $T(F)$  satisfying

$$\eta(T(F), F) = \int_{-\infty}^{\infty} \psi((x - T(F)) / S(F)) dF(x) = 0.$$

Note that the parameter of interest is  $T(F_0)$  and in general we have  $T(F) \neq T(F_0)$ . Hence, we need to bound the absolute difference

$$D(F) = |T(F) - T(F_0)| \quad (3)$$

between the M-location  $T(F)$  of the contaminated distribution and the M-location  $T(F_0) = \mu_0$  of the core (uncontaminated) distribution. The bias of robust estimates caused by outliers and other departures from symmetry can be assessed using the concept of maximum asymptotic bias. Martin and Zamar (1993) showed that for any  $F \in \mathcal{V}(F_0, \epsilon)$

$$|(T(F) - T(F_0)) / \sigma_0| \leq B(\epsilon),$$

where  $B(\epsilon)$  is the solution in  $t$  to  $\eta(t, \bar{F}) = 0$ , with  $\bar{F} = (1 - \epsilon)F + \epsilon\delta_x$ , where in general  $\delta_x$  is the point mass distribution at  $x$ . Therefore,  $D(F)$  is bounded by  $\sigma_0 B(\epsilon)$ . In general, however, this bound has to be improved to take into account the estimation of  $\sigma_0$ . This approach is presented in Section 2.

We now turn our attention to the problem of estimating the distribution of the robust estimate  $T(F_n)$ . In general, the finite sample distribution of robust estimates is unknown and inference will typically be based on their asymptotic distribution.

However, not much is known about the asymptotic distribution of robust estimates over the *entire* contamination neighborhood  $\mathcal{V}(F_0, \epsilon)$ .

When the distribution of the errors is symmetric the asymptotic distribution of the estimates is typically asymptotically normal and formulae for their asymptotic variance are available (see, for example, Hampel *et al.*, 1986). Unfortunately, in many contaminated data sets outliers appear to be asymmetrically distributed. If one relaxes the assumption of symmetry on the distribution of the errors, the calculation of the asymptotic variance of robust location and regression estimates becomes very involved (Carroll, 1978, 1979; Huber, 1981; Rocke and Downs, 1981; Carroll and Welsh, 1988; Salibian-Barrera, 2000). Salibian-Barrera (2000) shows that MM-location and regression estimates are asymptotically normal for any distribution in the contamination neighborhood. However, their asymptotic variance are difficult to estimate, since the formulae are numerically unstable. In Section 3 we describe a bootstrap method that is fast, stable and asymptotically correct for any  $F \in \mathcal{V}(F_0, \epsilon)$ .

The rest of the paper is organized as follows. In Section 2 we show how to find a computable bound for  $D(F)$  in (3) when  $\sigma_0$  is unknown. In Section 3 we review a computationally intensive method to estimate the distribution of robust estimates for arbitrary  $F \in \mathcal{V}(F_0, \epsilon)$ . Finally, in Section 4 we present globally robust confidence intervals for the simple linear model using the results of Sections 2 and 3.

## 2 Bias bounds

To compare competing robust estimates in terms of their bias behavior one can use the maxbias curve  $B(\epsilon)$  defined as

$$B(\epsilon) = \sup_{F \in \mathcal{V}(F_0, \epsilon)} \left| \frac{T(F) - T(F_0)}{\sigma_0} \right|$$

(the maximum of the standardized bias over a contamination neighborhood of size  $\epsilon$ ). Note that the above definition is an affine invariant quantity, and hence the actual values of the parameters do not affect the theoretical comparisons among the estimates. Two robustness measures closely related to the maxbias curve are the *contamination sensitivity* (CS) and the *breakdown point* (BP) introduced by Hampel (1974) and Hampel (1971), respectively. The CS gives a linear approximation for the maxbias curve

near zero. The BP is the largest fraction of contamination for which the maxbias curve remains bounded.

In this section, we show how to use the maxbias curve to find bias bounds (i.e. bounds for  $D(F)$  in (3)) for robust estimates in practical situations where we wish to construct robust confidence intervals or prediction intervals. The following example will show that in general to bound  $D(F)$  we need to take into account the values of the estimates of the nuisance parameters. To fix ideas, consider the location model and the median functional,  $M(F)$ . Suppose we have a large sample from a distribution  $F \in \mathcal{V}(F_0, \epsilon)$ . To assess the bias caused by outliers we must study the absolute difference  $|M(F) - M(F_0)|$  between the median of the contaminated distribution  $F$  and the central (uncontaminated) median of  $F_0$ . According to Huber's classical result (Huber, 1964) regarding the maxbias of the median, we have

$$\left| \frac{M(F) - M(F_0)}{\sigma_0} \right| \leq F_0^{-1} \left( \frac{1}{2(1 - \epsilon)} \right) = B_M(\epsilon)$$

and, therefore,  $|M(F) - M(F_0)| \leq \sigma_0 B_M(\epsilon)$ . However, in practice  $\sigma_0$  is seldom known and must be estimated by a robust scale functional  $S(F)$ , namely, the median of absolute deviations to the median (MAD). Unfortunately,  $S(F)B_M(\epsilon)$  is not an upper bound for  $|M(F) - M(F_0)|$  because  $S(F)$  may underestimate  $\sigma_0$ . For instance, if 10% of the data are outliers placed at 0.15, that is,  $F = 0.90N(0, 1) + 0.10\delta_{0.15}$ , then  $|M(F) - M(F_0)| = 0.1397 > S(F)B_M(0.10) = 0.8818 \times 0.1397 = 0.1232$ .

## 2.1 Bias bounds for location estimates

Recently, Berrendero and Zamar (2001) have introduced the concept of *bias bound* to tackle the above problem. A *bias bound* for  $T$  is a quantity  $K(\epsilon)$  such that

$$|T(F) - T(F_0)| \leq S(F)K(\epsilon), \tag{4}$$

for all  $F \in \mathcal{V}(F_0, \epsilon)$ . As shown in the example above, the maxbias curve *is not* in general a bias bound, although obviously both concepts are related. First, we will describe a quite straightforward method to obtain a bias bound in the location model. Since we cannot use the maxbias curve directly, due to the underestimation of the scale parameter, the idea is to consider the maximum conceivable underestimation.

Let  $S(F)$  be the auxiliary scale functional. Let  $S^-(\epsilon)$  be the implosion maxbias curve of the scale functional (see Martin and Zamar, 1993):

$$S^-(\epsilon) = \inf_{F \in \mathcal{V}_\epsilon(F_0)} \frac{S(F)}{\sigma_0}.$$

Then, for any location functional  $T(F)$ ,

$$|T(F) - T(F_0)| \leq \sigma_0 B_T(\epsilon) = S(F) \frac{\sigma_0}{S(F)} B_T(\epsilon) \leq S(F) \frac{B_T(\epsilon)}{S^-(\epsilon)}. \quad (5)$$

Therefore,  $K_1(\epsilon) \doteq B_T(\epsilon)/S^-(\epsilon)$  is a bias bound in the sense given by (4). We will call  $K_1(\epsilon)$  the *naive* bias bound. This bound can be sharpened using the following result (see Berrrendero and Zamar, 2001).

**Theorem 1** *Let  $T(F)$  be a location M-functional with score function  $\psi$  and auxiliary scale functional  $S(F)$ . That is,  $T(F)$  is defined as the solution of*

$$E_F \psi \left( \frac{X - T(F)}{S(F)} \right) = 0,$$

where  $\psi$  is continuous, increasing, odd and bounded with  $\psi(\infty) = 1$ . Denote  $S_0 = \inf_{F \in \mathcal{V}(F_0, \epsilon)} S(F)/\sigma_0$  and  $S_\infty = \sup_{F \in \mathcal{V}(F_0, \epsilon)} S(F)/\sigma_0$ . Then,

$$K_2(\epsilon) = \sup_{S_0 \leq s \leq S_\infty} \frac{\gamma(s)}{s},$$

is a bias bound for  $T$ , where  $\gamma(s)$  is implicitly defined by  $g[\gamma(s), s] = \epsilon/(1 - \epsilon)$ , with  $g(t, s) = -E_{F_0} \psi[(X - t)/s]$ .

In the case of the median, it is a simple exercise to show that both the naive and the improved bias bounds coincide, that is,  $K_1(\epsilon) = K_2(\epsilon)$ . However, for other location M-estimates, applying Theorem 1 may lead to a substantial improvement. We have computed  $B_T(\epsilon)$ ,  $K_1(\epsilon)$  and  $K_2(\epsilon)$  for several values of  $\epsilon$  and the Gaussian central model, when  $T(F)$  is a 95% efficient Huber's location M-estimate with score function  $\psi(x) = \min\{1, \max\{x/c, -1\}\}$ ,  $c = 1.345$  and scale estimate  $S(F) = \text{MAD}(F)$ . The results are displayed in Table 1.

Note that both bias bounds are greater than the maxbias since they take into account the underestimation of the scale. Moreover, whereas  $K_2(\epsilon)$  is not much greater

$\epsilon$	$B_T(\epsilon)$	$K_1(\epsilon)$	$K_2(\epsilon)$
0.05	0.09	0.09	0.09
0.10	0.20	0.22	0.20
0.15	0.33	0.41	0.33
0.20	0.50	0.69	0.51
0.25	0.74	1.16	0.76
0.30	1.10	2.02	1.13
0.35	1.67	3.83	1.79
0.40	2.63	8.44	3.19

Table 1: Maxbias and bias bounds for Huber’s location M-estimates (S=MAD) when  $F_0 = \Phi$  is the standard normal distribution.

than  $B_T(\epsilon)$  for any amount of contamination, the behavior of the naive bound  $K_1(\epsilon)$  is quite deficient when  $\epsilon$  is large. This suggests that underestimating  $\sigma_0$  could be important for large values of  $\epsilon$  ( $\epsilon > 0.15$ ).

Even though the bias bound given by the above theorem is a significant improvement over the more naive one  $K_1(\epsilon)$ , we could still consider the problem of computing the optimal bias bound, given by

$$K^*(\epsilon) = \sup_{F \in \mathcal{V}(F_0, \epsilon)} \frac{|T(F) - T(F_0)|}{S(F)}.$$

This problem is more difficult since now both numerator and denominator depend simultaneously on the contaminated distribution  $F \in \mathcal{V}(F_0, \epsilon)$ . As far as we know this problem remains open.

## 2.2 Bias bounds for regression estimates

Finding bias bounds for regression estimates is a more demanding task than for location estimates because in the regression model we have to deal with more nuisance parameters. To keep the analysis at a relatively simple technical level, we will only consider here the case of Gaussian regressors in the regression-through-the-origin linear model. Bias bounds valid under broader conditions (the presence of intercept in the model and non Gaussian regressors) can be found in Berrendero and Zamar (2001).

In the rest of this section we will assume that we have  $n$  independent observations satisfying

$$y_i = \boldsymbol{\theta}'_0 \mathbf{x}_i + \sigma_0 u_i, \quad 1 \leq i \leq n$$

where the independent errors,  $u_i$ , have standard normal distribution  $F_0$  and are independent of the regressors. We assume that the regressors  $\mathbf{x}_i$  are independent random vectors with common distribution  $G_0$ . The joint distribution of  $(y_i, \mathbf{x}_i)$  under this model is denoted  $H_0$ . To allow for a fraction  $\epsilon$  of contamination in the data we assume that the actual true distribution  $H$  of  $(y_i, \mathbf{x}_i)$  lies within a contamination neighborhood of  $H_0$ ,  $\mathcal{V}(H_0, \epsilon)$ . Denote by  $F_{H, \mathbf{T}(H)}$  the distribution of the residuals  $y_i - \mathbf{T}(H)' \mathbf{x}_i$  produced by a regression affine equivariant functional  $\mathbf{T}$  under  $H$ .

The maxbias curve of  $\mathbf{T}$  is again defined as an invariant quantity (see Martin *et al.* (1989)):

$$B_{\mathbf{T}}(\epsilon) = \sup_{H \in \mathcal{V}(H_0, \epsilon)} \{[\mathbf{T}(H) - \boldsymbol{\theta}_0]' \Sigma_0 [\mathbf{T}(H) - \boldsymbol{\theta}_0]\}^{1/2} / \sigma_0.$$

The matrix  $\Sigma_0$  is the covariance matrix of the regressors under  $G_0$  (although other affine equivariant scatter matrices could also be chosen.) Starting from this definition, it is possible to obtain an upper bound for the difference  $\|\mathbf{T}(H) - \mathbf{T}(H_0)\|$ , which depends on the nuisance parameters. In fact, it is not difficult to prove (see Lemma 2 in Berrendero and Zamar, 2001) that, for all  $H \in \mathcal{V}(H_0, \epsilon)$ ,

$$\|\mathbf{T}(H) - \mathbf{T}(H_0)\| \leq \frac{\sigma_0}{\sqrt{\lambda_1(G_0)}} B_{\mathbf{T}}(\epsilon),$$

where  $\lambda_1(G_0)$  is the smallest eigenvalue of  $\Sigma_0$ . Similarly to the location case (see equation (5) above) we must now estimate  $\sigma_0$  and  $\lambda_1(G_0)$  and consider the problems caused by the bias in the estimation of these parameters. The residual scale  $\sigma_0$  can be estimated using a scale functional applied to the distribution of the residuals  $S_1(F_{H, \mathbf{T}(H)})$ . On the other hand, the estimation of  $\lambda_1(G_0)$  poses an interesting problem that links the computation of bias bounds in the regression model with robust techniques of multivariate analysis. In particular, some sort of robust principal components analysis seems appropriate. We adopt here the projection pursuit approach proposed by Li and Chen (1985). It is well known the following property of the smallest eigenvalue:



$\lambda_1(G_0) = \min_{\|\mathbf{a}\|=1} \text{Var}_{G_0}(\mathbf{a}'\mathbf{x})$ . The idea is to replace the variance with a robust dispersion estimate of the projections  $\mathbf{a}'\mathbf{x}$  in the last formula. That is, we define

$$\hat{\lambda}_1(G) = \left[ \min_{\|\mathbf{a}\|=1} S_2(G, \mathbf{a}) \right]^2,$$

where  $S_2(G, \mathbf{a})$  is a robust scale of the projections  $\mathbf{a}'\mathbf{x}$  under  $G$ , where  $G$  is an arbitrary distribution function. With a similar argument to that of equation (5) we can now prove that

$$\|\mathbf{T}(H) - \mathbf{T}(H_0)\| \leq \frac{S_1(F_{H, \mathbf{T}(H)}) S_1^+(\epsilon)}{\sqrt{\hat{\lambda}_1(G)} S_2^-(\epsilon)} B_{\mathbf{T}}(\epsilon),$$

where  $S_1^+(\epsilon)$  and  $S_2^-(\epsilon)$  are, respectively, the explosion and implosion maxbias curves of the scales used to estimate  $\sigma_0$  and  $\lambda_1(G_0)$ . Hence,

$$K_{\mathbf{T}}(\epsilon) = \frac{S_1^+(\epsilon)}{S_2^-(\epsilon)} B_{\mathbf{T}}(\epsilon) \tag{6}$$

is a bias bound for  $\mathbf{T}$ .

### 3 Estimating asymptotic distributions and asymptotic variances of robust estimates

Recently some effort has been devoted to finding *global* asymptotic properties of robust estimates over neighborhoods of distributions (see Davies, 1993). Salibian-Barrera and Zamar (2001) showed that M-location estimates calculated with an S-scale (which, following Yohai (1987), we call *MM-location* estimates) are consistent to their asymptotic value, uniformly on the contamination neighborhood. Formally, for any  $\delta > 0$  we have

$$\lim_{m \rightarrow \infty} \sup_{F \in \mathcal{V}(\Phi, \epsilon)} P_F \left\{ \sup_{n \geq m} |T(F_n) - T(F)| > \delta \right\} = 0. \tag{7}$$

There is a trade-off between the size  $\epsilon$  of the contamination neighborhoods where (7) holds and the breakdown point of the scale estimate. Table 2 lists the maximum values

BP	0.50	0.45	0.40	0.35	0.30	0.25
$\epsilon$	0.11	0.14	0.17	0.20	0.24	0.25

Table 2: Maximum size  $\epsilon$  of contamination neighborhoods where uniform consistency of MM-location estimates holds for different breakdown points (BP).

of  $\epsilon$  such that (7) holds for MM-location estimates with breakdown points between 0.25 and 0.50 and contamination neighborhoods  $\mathcal{V}(\Phi, \epsilon)$  of the standard normal distribution  $\Phi$ . With additional regularity conditions these MM-location estimates are asymptotically normal for any  $F \in \mathcal{V}(\Phi, \epsilon)$ . Moreover, the weak convergence is uniform in  $\mathcal{V}(\Phi, \epsilon)$ :

$$\lim_{n \rightarrow \infty} \sup_{F \in \mathcal{V}(\Phi, \epsilon)} \sup_{x \in \mathbb{R}} \left| P_F \left\{ \sqrt{n} \left( \hat{\mu}_n - \mu(F) \right) < x \sqrt{V} \right\} - \Phi(x) \right| = 0. \quad (8)$$

where  $V = V(T(F), \sigma, F)$  is the asymptotic variance of  $T(F_n)$  (see Salibián-Barrera and Zamar, 2001). This last result is of practical importance since, for example, it guarantees that the sample size needed for a good asymptotic approximation to the distribution of  $T(F_n)$  does not depend on the distribution of the data.

Once we know the asymptotic distribution of robust estimates for arbitrary  $F \in \mathcal{V}(\Phi, \epsilon)$ , in order to perform statistical inference based on them we need to estimate their asymptotic variance. In general, consistent estimates of these variances can be obtained from the corresponding asymptotic formulae. For example, to estimate the asymptotic variance  $V$  of MM-location estimates in (8) we can use  $\hat{V} = V(T(F_n), \hat{\sigma}_n, F_n)$ . However, the involved form of  $V$  (and its matrix counterpart for linear regression) produces unstable estimates which in turn result in unsatisfactory confidence intervals.

### 3.1 Bootstrapping robust estimates

Another method to estimate the variability of statistical estimates is given by the bootstrap (Efron, 1979). This method has been extensively studied for diverse models. In particular, the theory for bootstrap distribution of robust estimates has been studied by Shorack (1982), Parr (1985), Yang (1985), Shao (1990, 1992), Liu and Singh (1992) and Singh (1998).

Two problems of practical relevance arise when bootstrapping robust regression estimates. First, the proportion of outliers in the bootstrap samples may be higher than that in the original data set causing the bootstrap quantiles to be very inaccurate. The second difficulty is caused by the heavy computational requirements of many robust estimates.

Intuitively, the reason for the first problem above is that with a certain positive probability the proportion of outliers in a bootstrap sample will be larger than the break-down point of the estimate. Thus, the tails of the bootstrap distribution may be affected by the outliers.

The high computational demand of robust estimates (specially in linear regression models) may render the method unfeasible for moderate to high dimensional problems. Moreover, due to the first problem mentioned above, even if we wait until the calculations are done, the resulting estimate may not be reliable (for example, the tails of the bootstrap distribution might be highly inaccurate).

The first problem (high proportion of outliers in the bootstrap samples) has been studied for location robust estimates by Singh (1998). To obtain consistent bootstrap quantile estimates with a high breakdown point Singh proposed to Winsorize the observations around the robust location estimates and then to re-sample from the Winsorized observations. Unfortunately it is not clear how to extend this method to linear regression models.

In recent years the feasibility of bootstrapping computationally demanding estimates (second problem above) has received some attention in the literature (Schucany and Wang, 1991; Hu and Kalbfleisch, 2000). Unfortunately, the regularity conditions needed for their proposal are not satisfied by robust regression estimates.

Salibian-Barrera and Zamar (2002) introduce a bootstrap method that simultaneously addresses both problems above. Namely, it is resistant to the presence of outliers in the data and it is computationally simple. The basic idea is best presented for the simple location model with known scale. Let  $x_1, \dots, x_n$  be a random sample satisfying

$$x_i = \mu + \epsilon_i, \quad i = 1, \dots, n, \quad (9)$$

where  $\epsilon_i$  are independent and identically distributed random variables with known variance. Let  $\psi$  be odd and bounded. The associated M-location estimate for  $\mu$  is

defined as the solution  $T(F_n)$  of

$$\sum_{i=1}^n \psi(x_i - T(F_n)) = 0. \quad (10)$$

It is easy to see that  $T(F_n)$  can also be expressed as a weighted average of the observations:

$$T(F_n) = \frac{\sum_{i=1}^n \omega_i x_i}{\sum_{i=1}^n \omega_i}, \quad (11)$$

where  $\omega_i = \psi(x_i - T(F_n)) / (x_i - T(F_n))$ .

Let  $x_1^*, \dots, x_n^*$  be a bootstrap sample of the data (i.e. a random sample taken from  $x_1, \dots, x_n$  with replacement). We can recalculate  $T(F_n)$  using equation (11):

$$T(F_n)^* = \frac{\sum_{i=1}^n \omega_i^* x_i^*}{\sum_{i=1}^n \omega_i^*}, \quad (12)$$

with  $\omega_i^* = \psi(x_i^* - T(F_n)) / (x_i^* - T(F_n))$ . Note that we are not fully recalculating the estimate from each bootstrap sample, we only compute a weighted average (moreover, the weights do not have to be re-calculated either). Commonly used functions  $\psi$  yield weights  $\omega(u)$  that are decreasing functions of  $|u|$ . In this case, outlying observations that typically have a large residual will have a small weight in (11).

The re-calculated  $T(F_n)^*$ 's in (12) may not reflect the actual variability of  $T(F_n)$ . Intuitively this happens because the weights  $\omega_i$  in (11) are not re-computed with each bootstrap sample. To remedy this loss of variability we apply a correction factor  $a_n$  that does not need to be re-calculated. See Salibian-Barrera (2000) for a definition of  $a_n$ .

This method yields quantile estimates with high breakdown point. To obtain a consistent estimate of the asymptotic distribution of  $T(F_n)$  for any  $F \in \mathcal{V}(\Phi, \epsilon)$  we need to include a scale estimate in (10), (11) and in our re-calculations (12). We refer the interested reader to Salibian-Barrera and Zamar (2002) for a detailed discussion of the method in the linear regression context.

## 3.2 Applications and future directions

A direct application of the ‘‘robust bootstrap’’ discussed above is the construction of confidence intervals and tests for each parameter of the linear model (Salibian-

Barrera and Zamar, 2002). The problem of testing more general hypotheses on the vector of regression parameters using the robust bootstrap is studied in Salibian-Barrera (2002b). Consider the classes of robust tests discussed in Markatou *et al.* (1991) (see also Markatou and Hettmansperger, 1990). The distribution of these tests under the null hypothesis is known only for symmetric errors. It is of interest to be able to estimate the distribution of these tests under the null hypothesis in more general cases. By noting that these robust tests are functions of the robust estimate calculated under each hypothesis, the basic idea is to adapt the robust bootstrap to re-calculate the robust estimate under each hypothesis when the data is bootstrapped following the null model.

Another method to obtain fast estimates of the distribution of robust estimates is to bootstrap a one-step Newton-Raphson version of them. To fix ideas consider the simple location-scale model (9) above. Let  $T(F_n)$  be an MM-location estimate, let  $S(F_n)$  be the S-scale estimate and let  $\tilde{T}(F_n)$  be the associated S-location estimate. The system of equations

$$\begin{aligned} \sum_{i=1}^n \psi((x_i - T(F_n))/S(F_n)) &= 0, \\ \frac{1}{n} \sum_{i=1}^n \left[ \chi\left(\frac{x_i - \tilde{T}(F_n)}{S(F_n)}\right) - b \right] &= 0, \\ \frac{1}{n} \sum_{i=1}^n \chi'\left(\frac{x_i - \tilde{T}(F_n)}{S(F_n)}\right) &= 0. \end{aligned}$$

can be written as

$$\mathbf{g}_n(\hat{\boldsymbol{\theta}}_n) = 0,$$

where  $\mathbf{g}_n : \mathfrak{R}^3 \rightarrow \mathfrak{R}^3$  and  $\boldsymbol{\theta}_n = (T(F_n), S(F_n), \tilde{T}(F_n))'$ . The Newton-Raphson iterations are

$$\boldsymbol{\theta}_{j+1} = \boldsymbol{\theta}_j - [\nabla \mathbf{g}_n(\boldsymbol{\theta}_j)]^{-1} \mathbf{g}_n(\boldsymbol{\theta}_j),$$

with  $\nabla \mathbf{g}_n(\boldsymbol{\theta}_j)$  the matrix of first partial derivatives of  $\mathbf{g}_n$  evaluated at  $\boldsymbol{\theta}_j$ . It is easy to see that for the MM-location estimate we obtain

$$\tilde{T}(F_n)^* = T(F_n) + S(F_n) \left[ \frac{\overline{\psi}(u)^*}{\overline{\psi}'(u)^*} + \frac{\overline{\psi}(u) u^*}{\overline{\psi}'(u)^*} \times \frac{\overline{\chi}(u) - b^*}{\overline{\chi}'(u) u^*} \right],$$

where  $\overline{\psi}(u)^* = \overline{\psi(u)^*} = \sum \psi((x_i^* - T(F_n))/S(F_n))/n$  and similarly the others. In Salibian-Barrera (2002a) it is shown that the above procedure yields a consistent estimate of the asymptotic distribution of  $T(F_n)$ . However, note that the numerically unstable denominators above make the method susceptible to aberrant values when the bootstrap sample contains many outliers. On the other hand, the fact that this method does not need a correction factor might produce a theoretically better approximation to the distribution of  $T(F_n)$  (that is, it could inherit the  $O_p(1/n)$  order that is expected from most bootstrap methods). Some of these questions are addressed in Salibian-Barrera (2002b).

## 4 Globally Robust Confidence Intervals

It is well known that given a random sample  $x_1, \dots, x_n$  drawn from a normal population, that is

$$x_i \sim N(\mu, \sigma^2), i = 1, \dots, n \quad (13)$$

a level- $(1 - \alpha)$  confidence interval for  $\mu$  is given by

$$P_\mu \left( \mu \in \left[ \bar{x}_n - t_{n-1, \alpha/2} \frac{S_n}{\sqrt{n}}, \bar{x}_n + t_{n-1, \alpha/2} \frac{S_n}{\sqrt{n}} \right] \right) = 1 - \alpha,$$

where  $\bar{x}_n$ ,  $t_{n-1, \delta}$  and  $S_n$  stand for the mean average, the  $\delta$ -percentile of the Student- $t$  distribution and the sample standard deviation respectively. Then, the length of the intervals is

$$L = 2t_{n-1, \alpha/2} \frac{S_n}{\sqrt{n}}.$$

In this case, *coverage* and *length* depend on the assumption of normality. It seems rather natural to wonder whether this classical confidence interval may be *defective* when (13) is *only approximately valid*. More precisely, we have introduced earlier  $\epsilon$ -contamination neighborhoods  $\mathcal{V}(F, \epsilon)$  (see equation (1)) to represent the idea of having the majority of data points coming from the *parametric model* and the remaining ones coming from an *unknown distribution*. The purpose of robust methods is to

safeguard against deviations from the assumptions. Then, robustness is also concerned with the way coverage and length of a confidence interval are affected when the true underlying distribution is a member of the  $\epsilon$ -contamination neighborhood rather than the parametric model. Furthermore, the crucial issue is to show what alternative inferential procedures can be implemented to overcome somehow the damage caused by the departures from the parametric model. The vast majority of the robustness literature focuses on point estimation. For recent papers on robust inference see, for example, Markatou *et al.* (1991) and references therein.

The following example taken from Adrover *et al.* (2002) will show the effect of an outlier generating distribution on coverage and length for the Student- $t$  confidence interval. To see that, let us consider first the contaminated distribution

$$F^{x_0} = (1 - \epsilon)F_0 + \epsilon\delta_{x_0},$$

where  $x_0 > 0$ . Then,

$$L_n = \bar{x}_n - t_{n-1, \alpha/2} S_n / \sqrt{n} \rightarrow \epsilon x_0 \leftarrow \bar{x}_n + t_{n-1, \alpha/2} S_n / \sqrt{n} = U_n$$

as  $n$  tends to infinity, for all  $x_0$ . Therefore,

$$\lim_{n \rightarrow \infty} P_{F^{x_0}}(L_n < 0 < U_n) = 0,$$

which means that the interval is shrinking to  $\epsilon x_0$  and the coverage of the interval is tending to 0 as  $n$  tends to  $\infty$ . The asymmetry of the distribution causes the estimate to be *biased* and the interval fails to cover the true parameter 0.

Let us take now  $F^*$  a point mass distribution at  $\pm x_0$  (equally weighted). Since the standard deviation  $S_n$  is highly sensitive to the extreme observation  $x_0$ , the length of the interval becomes unbounded as  $x_0$  tends to  $\infty$ ,

$$\lim_{n \rightarrow \infty} \sup_{x_0 > 0} 2 t_{n-1, \alpha/2} \frac{S_n}{\sqrt{n}} \geq \lim_{n \rightarrow \infty} \sup_{x_0 > 0} 2 t_{n-1, \alpha/2} \frac{x_0}{\sqrt{n}} = \infty$$

Summing up, coverage and length are spoiled by the presence of outliers. Therefore, it seems natural to require *robust confidence intervals* to have the guaranteed probability coverage of the target parameter for all the distributions on the contaminated neighborhood as well as a reasonable average length uniformly over the entire neighborhood. Both desirable features have been referred in the literature as the *robustness*

of *validity* and *robustness of efficiency* of confidence intervals respectively (see Tukey and McLaughlin (1963), Dixon and Tukey (1968), Huber (1968, 1970), Barnett and Lewis, 1994, p. 74, and references therein, Fraiman *et al.* (2001) and Adrover *et al.* (2002)). Adrover *et al.* (2002) also deal with similar concepts of *stability* and *informativeness* which roughly parallel the robustness of validity and efficiency respectively. More precisely, they define

**Definition 1** A confidence interval  $(L_n, U_n)$  for  $\mu$  is called globally robust of level  $1 - \alpha$  if it satisfies the following conditions:

1. (Stable interval) The minimum asymptotic coverage over the  $\epsilon$ -contamination neighborhood is  $(1 - \alpha)$ :

$$\lim_{n \rightarrow \infty} \inf_{F \in \mathcal{V}(F_\mu, \epsilon)} P_F(L_n < \mu < U_n) \geq (1 - \alpha);$$

2. (Informative interval) The maximum asymptotic length of the interval is bounded over the  $\epsilon$ -contamination neighborhood:

$$\lim_{n \rightarrow \infty} \sup_{F \in \mathcal{V}(F_\mu, \epsilon)} [U_n - L_n] < \infty.$$

Mainly, there have been two approaches to overcome the lack of robustness of the classical confidence intervals. The *first approach* relies on the idea of replacing the mean average  $\bar{x}_n$  by a robust asymptotically normal location estimate  $T_n$ , and  $S_n/\sqrt{n}$  by an appropriate robust estimate of the standard error of  $T_n$ . If the estimate  $\hat{\sigma}_n$  of the standard error of  $T_n$  is uniformly bounded over the entire neighborhood, such procedure is successful in achieving the *robustness of efficiency*. Then, the confidence interval is given by

$$[T_n - z_{\alpha/2}\hat{\sigma}_n, T_n + z_{\alpha/2}\hat{\sigma}_n] \tag{14}$$

However, it still fails to get the nominal level. This is due to the fact that we can find distributions in the neighborhood such that  $T_n \rightarrow \tilde{\mu} \neq \mu$ , and then,

$$L_n = T_n - z_{\alpha/2}\hat{\sigma}_n \rightarrow \tilde{\mu} \leftarrow T_n + z_{\alpha/2}\hat{\sigma}_n = U_n.$$



Then, the confidence interval shrinks to  $\tilde{\mu}$  failing to achieve the nominal level. This entails inexorably that the goal of validity seems rather difficult to be achieved if the crucial issue of the asymptotic bias is neglected.

Frailman *et al.* (2001) and Adrover *et al.* (2002) provide a small Monte Carlo simulation which illustrates this point. Ten thousand standard normal samples of different sizes were generated, containing various fractions of contamination. The contaminating distribution is a point-mass distribution at  $x_0 = 4.0$ . Similar results were found for other asymmetric outlier generating distributions. For each sample, the location M-estimate was calculated with Huber  $\psi$ -function [ $\psi(y) = \min\{-c, \max\{c, y\}\}$ ], with truncation constant  $c = 1.345$ , and the corresponding asymptotic 95% confidence interval based on the empirical asymptotic variance. The coverage and average length of these intervals are given on Table 3.

Huber (1968) also noticed this phenomenon and broke new ground by establishing a remarkable finite sample optimality result. He considered the pure location model (9) (known scale) and intervals of fixed length  $2a$ . He minimized the quantity

$$\max_{F \in \mathcal{V}(F_0, \epsilon)} \max\{ P_F(\mu < \hat{\mu}_n - a) , P_F(\mu > \hat{\mu}_n + a) \}$$

over the class of location M-estimators  $\hat{\mu}_n$ . In principle, the value of  $a$  could be varied to obtain the desired maximum level for each  $n$  and  $\epsilon$ .

The actual problem of bias behind the lack of coverage in confidence intervals has been highlighted. Then, the *second approach* to accomplish robust confidence intervals according to Definition 1 takes into account a bias correction to generate stable coverage over the neighborhood. In that direction we can cite so far the papers by Frailman *et al.* (2001) in the context of the location model, Adrover *et al.* (2002) for the simple linear regression model and Yohai and Zamar (2001) in a non-parametric framework.

We briefly sketch the procedure to get the robust intervals. Let  $x_1, \dots, x_n$  be a random sample whose *unknown* distribution  $F$  belongs to the  $\epsilon$ -contaminated neighborhood. The proposal relies on three elements:

1. A robust asymptotically normal location estimate  $\hat{\mu}_n$ , that is

$$\sqrt{n}(\hat{\mu}_n - \hat{\mu}(F)) \rightarrow N(0, \sigma^2(F))$$

$\epsilon$	$n$	% of Coverage	Average Length
0.05	20	92%	0.91
	50	92%	0.60
	100	88%	0.44
	200	82%	0.31
0.10	20	91%	1.05
	50	84%	0.68
	100	67%	0.49
	200	39%	0.35
0.15	20	88%	1.19
	50	72%	0.76
	100	35%	0.56
	200	5%	0.40
0.20	20	82%	1.41
	50	45%	0.92
	100	8%	0.66
	200	0%	0.47

Table 3: Percentage of coverage and average length for 10000 asymptotic 95% confidence intervals based on Huber's location M-estimate.

2. A *known bias bound*  $\bar{\mu}$  such that

$$|\hat{\mu}(F) - \mu| \leq \bar{\mu}, \text{ for all } F \text{ in the neighborhood.}$$

3. The estimation of the asymptotic variance  $\sigma^2(F)$ .

Regarding the asymptotically normal robust estimates mentioned in 1., Fraiman *et al.* (2001) dealt with the class of M-estimators of location, finding the optimal location estimate in the sense of minimizing the maximum length of the confidence intervals. Adrover *et al.* (2002) considered a class of estimators in the simple linear regression model, called *median-based estimators*. This class of estimators was taken into account because of its remarkable asymptotic bias performance (see Adrover and Zamar (2000)). In this class are included the median of pairwise slopes (Theil (1950) and Sen (1968)), the repeated median of slopes (Siegel (1982)), the Brown and Mood's estimate (1951). Estimates with small bias bounds are needed to produce relatively short confidence intervals of practical relevance. Since they are interested in linear combinations of the regression coefficients, they need bounds on the biases of the slope and intercept parameters.

Even though the quantity  $b(F) = \hat{\mu}(F) - \mu$  (*bias*) is *unknown*, a *known bias bound*  $\bar{\mu}$  for  $b$  is available (see Section 2).

The scheme for constructing the robust confidence intervals is as follows. If we knew the distribution of  $\hat{\mu}_n$ , a level- $(1 - \alpha)$  confidence interval for  $\mu$  would be given by

$$(\hat{\mu}_n - l_n(F), \hat{\mu}_n + r_n(F))$$

where

$$P_F(-r_n(F) \leq \hat{\mu}_n - \mu \leq l_n(F)) = 1 - \alpha.$$

But  $l_n = l_n(F)$  and  $r_n = r_n(F)$  are *unknown* since  $F$  is assumed to be *partially known* in the robust setup. Since  $\hat{\mu}_n$  is asymptotically normal, we would be able to approximate  $l_n$  and  $r_n$  if we knew  $b$ ,

$$\Phi\left(\frac{l_n - b}{\hat{\sigma}_n}\right) + \Phi\left(\frac{r_n + b}{\hat{\sigma}_n}\right) - 1 = 1 - \alpha \quad (15)$$

where  $\sqrt{n}\hat{\sigma}_n$  is a consistent robust estimate for  $\sigma(F)$ . But the equation is still dependent on the *unknown* quantity  $b$ . Two different situations may be taken into account according to the degree of uncertainty:

1. It is known that the center of symmetry is shifted away either to the left or to the right (*bias constraint*),
2. There is no information on the sign of the bias (*unconstrained bias*).

In case 1., the equation (15) changes so as to reflect that the confidence interval must be shifted to the left in case of positive bias or viceversa in case of negative bias to keep the nominal level, that is,

$$\Phi\left(\frac{l_n - b}{\hat{\sigma}_n}\right) + \Phi\left(\frac{r_n}{\hat{\sigma}_n}\right) - 1 = 1 - \alpha \quad \text{when } b \geq 0,$$

$$\Phi\left(\frac{l_n}{\hat{\sigma}_n}\right) + \Phi\left(\frac{r_n + b}{\hat{\sigma}_n}\right) - 1 = 1 - \alpha \quad \text{when } b \leq 0.$$

After some manipulation, the intervals with minimal length in both situations turn out to be

$$\hat{I}_n(F) = (\hat{\mu}_n - \hat{\sigma}_n z_{\alpha/2} - b, \hat{\mu}_n + \hat{\sigma}_n z_{\alpha/2}) \quad \text{when } b \geq 0,$$

$$\hat{I}_n(F) = (\hat{\mu}_n - \hat{\sigma}_n z_{\alpha/2}, \hat{\mu}_n + \hat{\sigma}_n z_{\alpha/2} - b) \quad \text{when } b \leq 0.$$

But the interval  $\hat{I}_n(F)$  still depends on the unknown  $F$  and  $b$ . Then, to prevent from any  $F$  in the neighborhood causing bias, we can take

$$\begin{aligned} I_n &= \bigcup_{F \in \mathcal{V}(F_\mu, \epsilon)} (\hat{\mu}_n - \hat{\sigma}_n z_{\alpha/2} - b, \hat{\mu}_n + \hat{\sigma}_n z_{\alpha/2}) \\ &= (\hat{\mu}_n - \hat{\sigma}_n z_{\alpha/2} - \bar{\mu}, \hat{\mu}_n + \hat{\sigma}_n z_{\alpha/2}), \quad \text{when } b \geq 0. \end{aligned} \quad (16)$$

Similarly,

$$I_n = (\hat{\mu}_n - \hat{\sigma}_n z_{\alpha/2}, \hat{\mu}_n + \hat{\sigma}_n z_{\alpha/2} + \bar{\mu}), \quad \text{when } b \leq 0.$$

In case 2. there is no prior information on the bias and the search for the shortest interval from (15) leads to

$$\hat{I}_n(F) = (\hat{\mu}_n - \hat{\sigma}_n z_{\alpha/2}, \hat{\mu}_n + \hat{\sigma}_n z_{\alpha/2}) - b.$$

Observe that  $\hat{I}_n(F)$  is a shift of the interval (14) so as to cover the parameter  $\mu$ . Since  $\hat{I}_n(F)$  still depends on the unknown  $F$  and  $b$ , a robust interval of minimum level  $1 - \alpha$  over the neighborhood is defined as

$$I_n = (\hat{\mu}_n - \hat{\sigma}_n z_{\alpha/2} - \bar{\mu}, \hat{\mu}_n + \hat{\sigma}_n z_{\alpha/2} + \bar{\mu}). \quad (17)$$

A shorter robust confidence interval may be obtained by restricting the search to centered intervals about  $\hat{\mu}_n$  rather than  $\hat{I}_n(F)$ . The idea is to first consider intervals of the form

$$\hat{\mu}_n \pm q_n,$$

where  $q_n$  is the *true*  $(1 - \alpha)$ -quantile, that is,

$$P_F (|\hat{\mu}_n - \mu| \leq q_n) = 1 - \alpha. \quad (18)$$

Using (15) we can approximate (18) by

$$\Phi \left( \frac{\tilde{q}_n - b}{\hat{\sigma}_n} \right) + \Phi \left( \frac{\tilde{q}_n + b}{\hat{\sigma}_n} \right) - 1 = 1 - \alpha, \quad (19)$$

which yields the confidence interval

$$\hat{\mu}_n \pm \tilde{q}_n. \quad (20)$$

Fraiman *et al.* (2001) showed that  $\sqrt{n}(\tilde{q}_n - q_n) = o_p(1)$ . The solution  $\tilde{q}_n$  to (19) is a strictly increasing function of  $|b|$  for each fixed  $\hat{\sigma}_n$  (and of  $\hat{\sigma}_n$  for each fixed  $|b|$ ). That is, the quantile  $\tilde{q}_n$  is a monotone function of the bias (and standard deviation) and its largest possible value is obtained by replacing  $|b|$  by  $\bar{\mu}$ :

$$\Phi \left( \frac{\bar{q}_n - \bar{\mu}}{\hat{\sigma}_n} \right) + \Phi \left( \frac{\bar{q}_n + \bar{\mu}}{\hat{\sigma}_n} \right) - 1 = 1 - \alpha. \quad (21)$$

This yields the robust confidence interval

$$\hat{\mu}_n \pm \bar{q}_n. \quad (22)$$

This robust interval is shorter (more informative) than (17), that is,  $\hat{\sigma}_n z_{\alpha/2} + \bar{\mu} > \bar{q}_n$ .

	Degree of Knowledge		
	Unknown Bias Sign	Positive Bias	Negative Bias
LCB	$(\hat{\mu}_n - \hat{\sigma}_n z_\alpha - \bar{\mu}, \infty)$	$(\hat{\mu}_n - \hat{\sigma}_n z_\alpha - \bar{\mu}, \infty)$	$(\hat{\mu}_n - \hat{\sigma}_n z_\alpha, \infty)$
UCB	$(-\infty, \hat{\mu}_n + \hat{\sigma}_n z_\alpha + \bar{\mu})$	$(-\infty, \hat{\mu}_n + \hat{\sigma}_n z_\alpha)$	$(-\infty, \hat{\mu}_n + \hat{\sigma}_n z_\alpha + \bar{\mu})$
CI	$\hat{\mu}_n \pm \bar{q}_n$ (see (21))	$(\hat{\mu}_n - \hat{\sigma}_n z_{\alpha/2} - \bar{\mu}, \hat{\mu}_n + \hat{\sigma}_n z_{\alpha/2})$	$(\hat{\mu}_n - \hat{\sigma}_n z_{\alpha/2}, \hat{\mu}_n + \hat{\sigma}_n z_{\alpha/2} + \bar{\mu})$

Table 4: Upper and lower robust confidence bounds and confidence intervals. LCB (UCB) stands for Lower (Upper) Confidence Bound and CI stands for Confidence Interval.

Since  $\bar{q}_n \geq \tilde{q}_n$ , the robust confidence interval turns out to be larger than strictly necessary for most distributions in the neighborhood. However, the enlargement of the confidence interval is the smallest one which allows us to achieve the nominal coverage over the entire neighborhood, provided that the upper bound  $\bar{\mu}$  is sharp. Observe that  $\bar{q}_n$  and  $\hat{\sigma}_n z_{\alpha/2} + \bar{\mu}$  converge to  $\bar{\mu}$  as  $n \rightarrow \infty$ , so that the asymptotic lengths of these two robust confidence intervals are minimal, provided that the upper bound  $\bar{\mu}$  is sharp. The derivation of lower and upper confidence bounds follows along the same lines. Table 4 summarizes the findings of robust confidence bounds and intervals.

Classical p-values are also affected when the parametric model is only approximately valid and the level of the tests is upset by the presence of outliers. This is a closely related problem to robust confidence intervals and bounds. The construction of robust p-values parallels the procedure used in the classical theory for hypothesis tests by inverting LCB, UCB and CI. Given a hypothesis problem for the parameter  $\mu$  whose alternatives are  $H_{a,1} : \mu > \mu_0$ ,  $H_{a,2} : \mu < \mu_0$  and  $H_{a,3} : \mu \neq \mu_0$ , for a given parameter  $\mu_0$ , the rejection rule for a level- $\alpha$  test is given by

$$\begin{aligned} \text{Reject } H_0 \text{ if } & \mu_0 \notin (\hat{\mu}_n(\alpha, F), \infty) \quad (H_{a,1}), \\ & \mu_0 \notin (-\infty, \hat{\mu}_n(\alpha, F)) \quad (H_{a,2}), \text{ and} \\ & \mu_0 \notin (\hat{\mu}_n - \tilde{q}_n(\alpha, F), \hat{\mu}_n + \tilde{q}_n(\alpha, F)) \quad (H_{a,3}) \end{aligned}$$

respectively.  $\tilde{q}_n = \tilde{q}_n(\alpha, F)$  represents a solution to (19) and  $\hat{\mu}_n(\alpha, F) = \hat{\mu}_n - \hat{\sigma}_n z_\alpha - b$  or  $\hat{\mu}_n(\alpha, F) = \hat{\mu}_n + \hat{\sigma}_n z_\alpha - b$  stands for either the LCB or the UCB respectively. The

Alternative Hypothesis	Degree of Knowledge		
	Unconstrained Bias	Positive Bias	Negative Bias
$H_1 : \mu > \mu_0$	$1 - \Phi\left(\frac{\hat{\mu}_n - \mu_0 - \bar{\mu}}{\hat{\sigma}_n}\right)$	$1 - \Phi\left(\frac{\hat{\mu}_n - \mu_0 - \bar{\mu}}{\hat{\sigma}_n}\right)$	$1 - \Phi\left(\frac{\hat{\mu}_n - \mu_0}{\hat{\sigma}_n}\right)$
$H_1 : \mu < \mu_0$	$\Phi\left(\frac{\hat{\mu}_n - \mu_0 + \bar{\mu}}{\hat{\sigma}_n}\right)$	$\Phi\left(\frac{\hat{\mu}_n - \mu_0}{\hat{\sigma}_n}\right)$	$\Phi\left(\frac{\hat{\mu}_n - \mu_0 + \bar{\mu}}{\hat{\sigma}_n}\right)$
$H_1 : \mu \neq \mu_0$	$2 - \Phi\left(\frac{ \hat{\mu}_n - \mu_0  - \bar{\mu}}{\hat{\sigma}_n}\right) - \Phi\left(\frac{ \hat{\mu}_n - \mu_0  + \bar{\mu}}{\hat{\sigma}_n}\right)$ .		

Table 5: Robust p-values for different testing situations and bias constraints.

rejection rule still depends on the unknown value  $b$ .

The  $p$ -values are customarily defined as

$$\hat{p}_n(F) = \inf \{ \alpha : \mu_0 \notin (\hat{\mu}_n(\alpha, F), \infty) \} = 1 - \Phi\left(\frac{\hat{\mu}_n - \mu_0 - b}{\hat{\sigma}_n}\right) \quad (H_{a,1})$$

$$\hat{p}_n(F) = \inf \{ \alpha : \mu_0 \notin (-\infty, \hat{\mu}_n(\alpha, F)) \} = \Phi\left(\frac{\hat{\mu}_n - \mu_0 - b}{\hat{\sigma}_n}\right) \quad (H_{a,2})$$

$$\hat{p}_n(F) = \inf \{ \alpha : \mu_0 \leq \hat{\mu}_n - \tilde{q}_n(\alpha, F) \} \text{ if } \mu_0 \leq \hat{\mu}_n \quad (H_{a,3})$$

$$\hat{p}_n(F) = \inf \{ \alpha : \hat{\mu}_n + \tilde{q}_n(\alpha, F) \leq \mu_0 \} \text{ if } \hat{\mu}_n \leq \mu_0 \quad (H_{a,3})$$

The last two cases entail the  $p$ -value

$$\hat{p}_n(F) = 2 - \Phi\left(\frac{|\hat{\mu}_n - \mu_0| - b}{\hat{\sigma}_n}\right) - \Phi\left(\frac{|\hat{\mu}_n - \mu_0| + b}{\hat{\sigma}_n}\right) \quad (H_{a,3}).$$

In the three cases, we can get rid of the unknown bias  $b$  by using the known bias bound  $\bar{\mu}$ . To ensure a global minimum level  $\alpha$  over the neighborhood we take the *robust p-value* as

$$\hat{p}_n^R = \sup \hat{p}_n(F),$$

where the range of the supremum depends on the degree of uncertainty. Formulas for  $\hat{p}_n^R$  obtained by applying the sup over the appropriate ranges are given on Table 5.

So far, the robust estimate and the known bias bound have been at the very core of the derivation of confidence intervals and  $p$ -values. It has been aforementioned that in the simple linear regression model,

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

it seems appropriate to use any member of the class called *median-based estimates*. This class of estimates exhibits much better bias performances compared to some other proposals such as MM- (Yohai, 1987) or  $\tau$ -estimates (Yohai and Zamar, 1988). Then, Adrover *et al* (2002) suggest using the estimates obtained as the solutions  $(\hat{\alpha}_n, \hat{\beta}_n)$  to the equations

$$0 = \frac{1}{n} \sum_{i=1}^n \text{sign} \left( y_i - \hat{\alpha}_n - \hat{\beta}_n(x_i - \hat{m}_n) \right) \text{sign}(x_i - \hat{m}_n), \quad (23)$$

$$0 = \frac{1}{n} \sum_{i=1}^n \text{sign}(y_i - \hat{\alpha}_n - \hat{\beta}_n(x_i - \hat{m}_n)). \quad (24)$$

where  $\hat{m}_n = \text{Median}\{x_i\}$ . The corresponding regression fit and the vertical line  $x = T(F_n)$  split the plane into four quarters containing equal number of points. This is precisely the defining property of Brown and Mood's estimate (1951). The bias behavior of these estimates has been analyzed in Adrover and Zamar (2000).

The third element mentioned in the construction of robust intervals was the estimation of variability. Adrover *et al.* (2002) conducted a simulation study to compare three different methods to estimate the asymptotic variability of the estimate for the slope parameter in the simple linear regression model. In the end, they recommend to estimate  $\sigma^2$  by using the shorth (see for instance Rousseeuw and Leroy, 1987) of the bootstrap distribution of the estimate, which appears to be a good compromise when they consider the achieved coverages and the median lengths.

To check the performance of the robust confidence intervals a Monte Carlo simulation was conducted using 1000 replicates of the following sampling situations: sample sizes  $n = 20, 40, 60, 80, 100$  and 200 from contaminated normal distributions  $(1 - \epsilon)N(\mathbf{0}, I) + \epsilon N(\boldsymbol{\eta}, \tau^2 I)$  with  $\boldsymbol{\eta}' = (\eta_x, \eta_y)$ ,  $\tau = 0.1$ ,  $\eta_x = 3$  and  $\eta_y = 1.5$  (2) for  $\epsilon = 0.05$  (0.10). This case is referred as the "mild contamination case". For the "medium contamination case",  $\eta_x = 5$  and  $\eta_y = 2.5$  for  $\epsilon = 0.05$  and  $\epsilon = 0.10$ . Finally, a "strong contamination case" was considered with  $\eta_x = 5$  and  $\eta_y = 15$  for  $\epsilon = 0.05$  and  $\epsilon = 0.10$ . The nominal confidence level in all the cases is 0.95. The empirical coverage and lengths of robust confidence intervals are summarized in Table 6.

The coverages reported in Table 6 are mostly below the nominal 95% level (except in the case of strong contamination with  $\epsilon = 0.10$ ). Further numerical analysis shows



$\epsilon$	$n$	Type of contamination		
		Mild	Medium	Strong
5%	20	0.94 (1.41)	0.94 (1.42)	0.91 (1.42)
	40	0.92 (1.01)	0.92 (1.01)	0.93 (1.04)
	60	0.92 (0.86)	0.92 (0.86)	0.94 (0.90)
	80	0.92 (0.76)	0.92 (0.77)	0.94 (0.79)
	100	0.92 (0.71)	0.92 (0.71)	0.94 (0.73)
	200	0.95 (0.57)	0.95 (0.58)	0.96 (0.58)
10%	20	0.95 (1.54)	0.95 (1.56)	0.95 (1.79)
	40	0.89 (1.17)	0.87 (1.18)	0.96 (1.43)
	60	0.87 (1.05)	0.85 (1.08)	0.97 (1.28)
	80	0.86 (1.00)	0.85 (1.04)	0.98 (1.18)
	100	0.87 (0.95)	0.86 (1.00)	0.98 (1.10)
	200	0.91 (0.83)	0.92 (0.89)	0.99 (0.95)

Table 6: Monte Carlo mean coverage and median length (in parenthesis) of robust confidence interval for the slope  $\beta$  using the shorth of the bootstrap distribution of  $\hat{\beta}_n$ .

that these few cases of overcoverage are associated with overestimation of the estimate's variability.

#### 4.1 Robust non-parametric inference for the median

Rieder (1981) addresses the problem of robustifying rank tests preserving their non-parametric nature. He considers one sided tests for the one and two sample problems, showing that the least favorable distribution under a given fraction of contamination does not depend on the target model. Yohai and Zamar (2001) construct non-parametric confidence intervals and two-sided tests for the median of the target distribution. This proposal overcomes the problem of lack of coverage without using a bias bound. They obtain the exact finite sample least favorable distribution (in the contaminated neighborhood) for the sign-test statistic. It turns out that the least favorable distribution is independent of the target distribution and therefore the robustified sign test and associated interval are simultaneously non-parametric and robust.

The robustified non-parametric interval for the unique median  $\mu$  of a continuous distribution  $F$  is constructed as follows. The classical non-parametric confidence interval based on the two sided sign test is given by

$$I_\alpha(\mathbf{x}_n) = [x_{(k+1)}, x_{(n-k)}]$$

where  $\mathbf{x}_n = (x_1, \dots, x_n)$  is a random sample drawn from  $F$ .  $k$  and  $n - k$  are determined so that  $P(k < Z_n < n - k) = 1 - \alpha$ , with  $Z_n = \sum_{i=1}^n I(x_i - \mu > 0) \sim \text{Bin}(n, p)$  and  $p = 1/2$ . The key point to robustify this interval relies on the fact that under contamination,  $Z_n$  is still binomial but the parameter  $p$  corresponds on the form of the contamination. By choosing  $p$  corresponding to the least favorable distribution on the neighborhood, the authors can ensure that

$$\inf_{G \in \mathcal{V}(F, \epsilon)} P_G(x_{(k+1)} < \mu < x_{(n-k)}) \geq 1 - \alpha$$

where  $P(k < Z_n < n - k) = 1 - \alpha$ , with  $Z_n \sim \text{Bin}(n, p)$  and  $p = (1 - \epsilon)/2$ .

## 5 References

- Adrover, J.G. and Zamar, R.H. (2000). Bias robustness of three median-based regression estimates. Technical Report No 194, Department of Statistics, University of British Columbia. Canada.
- Adrover, J.G., Salibian, M. and Zamar, R.H. (2002). Robust inference for the simple linear regression model. Submitted for publication.
- Barnett, B. and Lewis, T. (1994). *Outliers in statistical data*. Wiley & Sons. New York.
- Berrendero, J.R. and Zamar, R.H. (2001). Maximum bias curves for robust regression with non-elliptical regressors. *Ann. Statist.* **29** 224-251.
- Brown, G.W. and Mood, A. M. (1951). On median tests for linear hypotheses, in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Univ. of California Press, Berkeley, 159-166.
- Carroll, R. J. (1978). On almost sure expansion for M-estimates. *Ann. Statist.*, **6**, 314-318.

- Carroll, R.J. (1979). On estimating variances of robust estimators when the errors are asymmetric. *J. Amer. Statist. Assoc.* , **74**, 674-679.
- Carroll, R. J. and Welsh, A. H. (1988) A note on asymmetry and robustness in linear regression. *The American Statistician*, **42**, 285-287.
- Davies, P.L. (1993). Aspects of robust linear regression. *Ann. Statist.*, **21**, 1843-1899.
- Dixon, W. and Tukey, J. (1968). Approximate behavior of the distribution of Winsorized t (trimming/Winsorization). *Technometrics*, **10**, 83-98.
- Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1-26.
- Fraiman, R., Yohai, V.J. and Zamar, R.H. (2001). Optimal robust M-estimates of location. *Ann. Statist.* **29**, 194-223.
- Hampel, F.R. (1971). General qualitative definition of robustness. *Ann. Statist.* **42**, 1887-1896.
- Hampel, F.R. (1974). The influence curve and its role in robust estimation , *J. Amer. Statist. Assoc.*, **69**, 383-393.
- Hampel, F.R., Ronchetti, E.Z., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics. The approach based on influence functions*. New York: Wiley.
- Hu, F. and Kalbfleisch, J.D. (2000). The estimating function bootstrap. *The Canadian Journal of Statistics*, **28**, 449-499.
- Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, California.
- Huber, P.J. (1968). Robust confidence limits. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **10**, 269-278.
- Huber, P.J. (1970). Studentizing robust estimates, in: *Nonparametric Techniques in Statistical Inference*, M.L.Puri, Ed., Cambridge University Press, Cambridge, England.
- Huber, P.J. (1981). *Robust Statistics*. New York: Wiley.
- Li, G. and Chen, Z. (1985) Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Montecarlo. *J. Amer. Statist. Assoc.*, **80**, 759,766.

- Liu, R.Y. and Singh, K. (1992). Efficiency and robustness in resampling. *Ann. Statist.*, **20**, 370-384.
- Markatou, M. and Hettmansperger, T.P. (1990). Robust bounded-influence tests in linear models. *J. Amer. Statist. Assoc.*, **85**, 187-190.
- Markatou, M., Stahel, W.A., and Ronchetti, E. (1991). Robust M-type testing procedures for linear models. In *Directions in Robust Statistics and Diagnostics*, Part I, Stahel, W., Weisberg, S., eds. Springer-Verlag. pp. 201-220.
- Maronna, R.A. and Yohai, V.J. (1981). Asymptotic behavior of general  $M$ -estimates for regression and scale with random carriers. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **58**, 7-20.
- Martin, R. D., Yohai, V. J. and Zamar, R. H. (1989). Min-max bias robust regression. *Ann. Statist.*, **17**, 1608-1630.
- Martin, R.D. and Zamar, R. (1993) Bias robust estimation of scale. *Ann. Statist.* **21**, 991-1017.
- Parr, W.C. (1985). The bootstrap: some large sample theory and connections with robustness. *Statistics and Probability Letters*, **3**, 97-100.
- Rieder, H. (1981). Robustness of one- and two-sample rank tests against gross errors. *Ann. Statist.* **9** 245-265.
- Rocke, D.M. and Downs, G.W. (1981). Estimating the variances of robust estimators of location: influence curve, jackknife and bootstrap. *Communications in Statistics, Part B - Simulation and Computation*, **10**, 221-248.
- Rousseeuw, P.J and Leroy, A.M. (1987). *Robust regression and outliers detection*. Wiley, New York.
- Salibian-Barrera, M. (2000). Contributions to the Theory of Robust Inference. Unpublished Ph.D. Thesis. University of British Columbia, Department of Statistics, Vancouver, BC. Available on-line at  
<ftp://ftp.stat.ubc.ca/pub/matias/Thesis>.
- Salibian-Barrera, M. (2002a). Bootstrapping one-step Newton-Raphson iterations - accuracy versus numerical stability. Work in progress.
- Salibian-Barrera, M. (2002b). Robust bootstrap estimates for p-values of robust tests for the linear model. Work in progress.

- Salibian-Barrera, M. and Zamar, R.H. (2002). Bootstrapping robust estimates of regression, *Ann. Statist.*, **30**, No. 2, April 2002.
- Salibian-Barrera, M. and Zamar, R.H. (2001). Uniform asymptotics for robust location estimates when the scale is unknown, Submitted June 2001.
- Schucany, W.R. and Wang, S. (1991). One-step bootstrapping for smooth iterative procedures. *Journal of the Royal Statistical Society, Series B*, **53**, 587-596.
- Sen, P.K. (1968). Estimates of the regression coefficient based on Kendall's Tau. *J. Amer. Statist. Assoc.*, **63**, 1379-1389.
- Shao, J. (1990). Bootstrap estimation of the Asymptotic variances of statistical functionals. *Annals of the Institute of Statistical Mathematics*, **42**, 737-752.
- Shao, J. (1992). Bootstrap variance estimators with truncation. *Statistics and Probability Letters*, **15**, 95-101.
- Shorack, G.R. (1982). Bootstrapping robust regression. *Communications in Statistics, Part A - Theory and Methods*, **11**, 961-972.
- Siegel, A. (1982). Robust regression using repeated medians. *Biometrika*, **69**, 242-244.
- Singh, K. (1998). Breakdown theory for bootstrap quantiles. *Ann. Statist.*, **26**, 1719-1732.
- Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis, I, II, and III. In *Nederl. Akad. Wetensch. Proc.*, pp. 386-392, 521-525, and 1397-1412.
- Tukey, J. and McLaughlin, D. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization. *Sankhyā*, **A 25**, 331-352.
- Yang, S-S (1985). On bootstrapping a class of differentiable statistical functionals with applications to L- and M-estimates. *Statistica Neerlandica*, **39**, 375-385.
- Yohai, V.J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.*, **15**, 642-656.
- Yohai, V. J. and Zamar, R. (1988). High breakdown point estimates of regression by means of the minimization of an efficient scale. *J. Amer. Statist. Assoc.* **83** 406-413.

Yohai, V.J. and Zamar, R.H. (2001). Robust non-parametric inference for the median.  
Submitted for publication.