ROBUST BOOTSTRAP METHODS WITH AN APPLICATION TO GEOLOCATION IN HARSH LOS/NLOS ENVIRONMENTS

Stefan Vlaski [1] *Michael Muma* [2] *Abdelhak M. Zoubir* [2]

[1] Department of Electrical Engineering
University of California, Los Angeles
Box 951594, Los Angeles, CA 90095-1594, U.S.A.
Email: svlaski@ucla.edu

[2] Signal Processing Group
Technische Universität Darmstadt
Merckstraße 25, 64283 Darmstadt, Germany
Email: {muma, zoubir}@spg.tu-darmstadt.de

ABSTRACT

The bootstrap is a powerful computational tool for statistical inference that allows for the estimation of the distribution of an estimate without distributional assumptions on the underlying data, reliance on asymptotic results or theoretical derivations. On the other hand, robustness properties of the bootstrap in the presence of outliers are very poor, irrespective of the robustness of the underlying estimator. This motivates the need to robustify the bootstrap procedure itself. Improvements to two existing robust bootstrap methods are suggested and a novel approach for robustifying the bootstrap is introduced. The methods are compared in a simulation study and the proposed method is applied to robust geolocation.

Index Terms— bootstrap, robust, regression, geolocation

1. INTRODUCTION

Geolocation describes the task of locating a Mobile Terminal (MT) in a network of Fixed Terminals (FT) with known location. Outliers are present due to Non-Line-Of-Sight (NLOS) signal paths between MT and FT. Hence, robust estimates are required. Traditional estimators yield a point estimate, which, due to the stochastic nature of the process, is inherently stochastic. The bootstrap [1] allows for the estimation of the distribution of the estimate based on the original sample. It allows for the extraction of additional information from a given sample [2]. This is crucial in a practical setting, where a repetition of the experiment is impractical or impossible. However, the bootstrap is non-robust, irrespective of the robustness properties of the underlying estimator. Hence, robust bootstrap methods are required.

The bootstrap was first introduced by Efron in 1979 [3] as an alternative to the Jackknife. With increasing computational capabilities, a wide plethora of variations and applications emerged [4]. The lack of robustness of the classical bootstrap was recognized [5] and many robust bootstrap methods, such as the Stratified Bootstrap [6], the Influence Function Bootstrap [7] and the Fast and Robust Bootstrap [8] have been developed. Improvements to two of these methods are proposed. Furthermore, a new robust bootstrap method, the Robust Starting Point Bootstrap (RSPB), is presented and compared in an empirical simulations study. A plethora of methods for robust geolocation exist ([9], [10] among others). These yield point estimates. We demonstrate, how the RSPB can be utilized to extend these and other regression based methods to estimate the distribution of estimates.

This paper is organized as follows: Section 2 shows how geolocation can be described as a robust linear regression problem. Proposed improvements and a novel robust bootstrap method are described in Section 3. The application to robust geolocation and an empirical simulation study are presented in Sections 4 and 5 respectively.

2. PROBLEM FORMULATION

Geolocation describes the problem of locating a MT within a network of M FTs with known locations $(x_{FT,i}, y_{FT,i}), i = 1, ..., M$.

2.1. Robust Geolocation

The geolocation task can be formulated as a regression problem according to [9]

$$\boldsymbol{Y} = \boldsymbol{h}(x, y) + \boldsymbol{U}, \tag{1}$$

where $\boldsymbol{h}(x, y) = (h_1(x, y), ..., h_M(x, y))^T$ is a real, non-linear vector function describing the relation between the signal parameter used for the location estimation, the position of the MT and positions of FT's. For Time Of Arrival (TOA) estimation, it is the Euclidean distance between the MT and respective FT.

The above non-linear regression problem can be approximated with sufficient accuracy [9] as a linear regression problem [11], which gives

$$\boldsymbol{Y} = \boldsymbol{X}^T\boldsymbol{\beta} + \boldsymbol{U}, \tag{2}$$

where $\boldsymbol{X} \in \mathbb{R}^{p\times N}$ and $\boldsymbol{Y} \in \mathbb{R}^{N\times 1}$ are the predictor and response variables, respectively, and $\boldsymbol{U} \sim (1 - \varepsilon)\mathcal{N}(0, \sigma_e^2 I_N) + \varepsilon\mathcal{H}$ is the error term. In the application of geolocation, $\boldsymbol{U}$ is a superposition of low-variance noise,

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

which is also present in the LOS case, and high-mean, high-variance errors caused by NLOS signal paths with probability $\varepsilon = p_{nlos}$. $\boldsymbol{\beta}$ is the statistic of interest, i.e. the position of the MT.

2.2. The MM-Estimate of Regression

The linear approximation of Eq. (1) as Eq. (2) allows for the estimation of $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ through the MM-estimate [12]. It minimizes the loss function

$$\frac{1}{N}\sum_{i=1}^{N}\rho_2\left(\frac{r_i}{\hat{\sigma}}\right) = \delta. \tag{3}$$

Here, $r_i = y_i - \boldsymbol{x_i}^T\hat{\boldsymbol{\beta}}$ are residuals and $\hat{\sigma}$ is a scale S-estimate, minimizing the M-scale $\hat{\sigma}(\boldsymbol{\beta})$ defined by

$$\frac{1}{N}\sum_{i=1}^{N}\rho_1\left(\frac{r_i}{\hat{\sigma}(\boldsymbol{\beta})}\right) = b.$$

The constant b is adjusted to ensure consistency under the normal model. ρ_1 and ρ_2 are bounded ρ-functions [11]. ρ_1 determines the estimate's breakdown point, while ρ_2 determines efficiency. Hence, the MM-estimate can be tuned to be both robust and efficient [12]. The MM-estimate can be represented through an Iterative Re-weighted Least Squares (IRLS) fixed-point equation [13]

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} w_i(\hat{\boldsymbol{\beta}},\hat{\sigma},\boldsymbol{x_i},y_i)\boldsymbol{x_i}\boldsymbol{x_i^T}\right)^{-1}\sum_{i=1}^{N} w_i(\hat{\boldsymbol{\beta}},\hat{\sigma},\boldsymbol{x_i},y_i)\boldsymbol{x_i}y_i. \tag{4}$$

It is crucial in finding the optimal solution of Eq. (2) as the absolute minimum of Eq. (3).

3. PROPOSED METHODS

3.1. Improved Stratified Bootstrap

It is proposed in [6] to group the sample into a set number of stratas of equal length based on their residuals $r_i = y_i - \boldsymbol{x_i^T}\hat{\boldsymbol{\beta}}$. This groups potential outliers together and leads to more representative bootstrap samples in terms of the proportion of outliers. Two problems arise, however: the number of stratas offers a trade-off between robustness and consistency. Especially for small sample sizes, a large number of stratas results in significant underestimation of variability. Secondly, even a large number of stratas does not guarantee that every single bootstrap re-sample contains fractions of contamination below the breakdown point of the underlying estimator. Motivated by the two drawbacks of the stratified bootstrap in [6], the following stratification is proposed:

For an estimator of regression with C as the maximum number of contaminated samples before breakdown, obtain the stratified sample as follows:

- **Step 1.** Compute the estimate $\hat{\boldsymbol{\beta}}$ based on the original sample $\chi = (\boldsymbol{x_i}, y_i) : i = 1, ..., N$.
- **Step 2.** For every pair $(\boldsymbol{x_i}, y_i)$, compute the residual $r_i = y_i - \boldsymbol{x_i^T}\hat{\boldsymbol{\beta}}$ and obtain the sorted sample χ_s, so that $|r_{s,1}| \leq |r_{s,2}| \leq ... \leq |r_{s,N}|$.
- **Step 3.** Obtain the stratas of χ by:

$$\chi_1 = (x_{s,1}, y_{s,1}), ..., (x_{s,N-C}, y_{s,N-C})$$
$$\chi_2 = (x_{s,N-C+1}, y_{s,N-C+1}), ..., (x_{s,N}, y_{s,N})$$

By incorporating the breakdown point of the underlying estimator and allowing stratas of unequal length, robustness is guaranteed, as long as the original sample contains at most C outliers. For the MM-estimate with breakdown point $\varepsilon_{BP} = 0.5$ and sample of size $N = 20$, the tuning constant is set to $C = \lceil \varepsilon_{BP} N \rceil - 1 = 9$. The resulting bootstrap procedure matches the breakdown point of the underlying estimator.

3.2. Improved Influence Function Bootstrap

Robust estimates are characterized by bounded influence functions [11]. The Influence Function Bootstrap (IFB) [7] applies this concept to the bootstrap by assigning re-sampling probabilities based on the Influence Function (IF). By doing so, potential outliers are assigned small probabilities of appearing in any bootstrap re-sample.

Assigning smaller re-sampling probabilities to outlying samples reduces the effective sample size. By sampling N times with replacement from a sample of effective size N_{eff}, variability is underestimated and consistency is not achieved. This motivates the following improvement to the IFB: Obtain the effective sample size as

$$N_{\text{eff}} = \left\lfloor \sum_{i=1}^{N} w_i \right\rfloor,$$

where $\lfloor \cdot \rfloor$ indicates the integer part and $0 \leq w_i \leq 1$ is the re-sampling weight determined by the IFB. Sample N_{eff} instead of N times with replacement.

3.3. Robust Starting Point Bootstrap (RSPB)

Robustly estimating $\hat{\boldsymbol{\beta}}$ in Eq. (2) translates into finding the absolute minimum of a non-convex loss function, which is determined by the estimator. Both the Fast-MM [12] and Fast-τ [14] algorithms utilize the fixed-point Eq. (4), which reduces the loss function in every iteration until convergence to a local minimum. In the sequel, the Robust Starting Point Bootstrap (RSPB) is proposed. It is based on the following observation: For every bootstrap sample χ^*, it is not desirable to find the absolute minimum of the loss function, but rather a local minimum close to $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$, the estimates based on χ.

Non-convex minimization problems can generally be divided in two steps:

- **Step 1.** Obtain a good candidate point or a number of candidate points.
- **Step 2.** Apply an iterative algorithm, which reduces the respective loss function until the estimate converges to a solution of Eq. (4).

In the case of the MM-estimate, **Step 1.** is performed by a robust, but inefficient S-estimator, while an efficient, but not necessarily as robust M-estimate solves **Step 2.** For the τ-estimate of regression, **Step 1.** is performed by random re-sampling, upon which a number of candidate points are step-wise iterated and eliminated in **Step 2.** Performing the full algorithm on every bootstrap sample is both computationally demanding and non-robust as some samples may contain a larger fraction of contamination than the original sample.

For the RSPB, **Step 1.** is replaced by **Step 1'.** for every bootstrap sample as follows:

- **Step 0.** Obtain $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ for the original sample χ through the full algorithm described above. This is done only once prior to bootstrapping.
- **Step 1'.** For every bootstrap re-sample χ^*, choose the estimate $\hat{\boldsymbol{\beta}}$ with scale estimate $\hat{\sigma}$ from χ as the only candidate point for the minimization problem.
- **Step 2.** Apply an iterative algorithm, which reduces the respective loss function until the estimate converges to a solution of Eq. (4) for χ^* to obtain $\hat{\boldsymbol{\beta}}^*$ and $\hat{\sigma}^*$.

The above alteration ensures, that for every bootstrap sample χ^*, the estimates $\hat{\boldsymbol{\beta}}^*$ and $\hat{\sigma}^*$ converge to a local minimum close to $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$, rather than the absolute minimum, which may be skewed due to over-contamination. This results in a robust bootstrap replicate $\hat{\boldsymbol{\beta}}^*$, even if χ^* is contaminated past the breakdown point of the underlying estimator.

4. RSPB FOR GEOLOCATION

The RSPB can be applied directly to the linearized regression based approach to geolocation in Eq. 2 and [9]. Fig. 1 displays the RSPB distribution estimate (contour plot) of the the MM-estimate, the MM-point estimate and the true position of the MT for the LOS case ($p_{\text{nlos}} = 0$). FT's are located at $(2500, 5000)$, $(1500, 4000)$, $(3000, 4500)$, $(4000, 3500)$, $(2000, 250)$ and $(1000, 1000)$.

Fig. 2 displays the RSPB distribution estimate in the mixed LOS/NLOS case with $p_{\text{nlos}} = 0.4$. The distribution estimate adapts to the MM-estimate, which, due to the strong contamination performs slightly worse than in the LOS case. Nonetheless, the quality of the distribution estimate is comparable to the LOS case. Furthermore, unlike the classical bootstrap, the RSPB distribution estimate inherits the robustness properties of the underlying MM-estimate and is similar in terms of area to the LOS case.

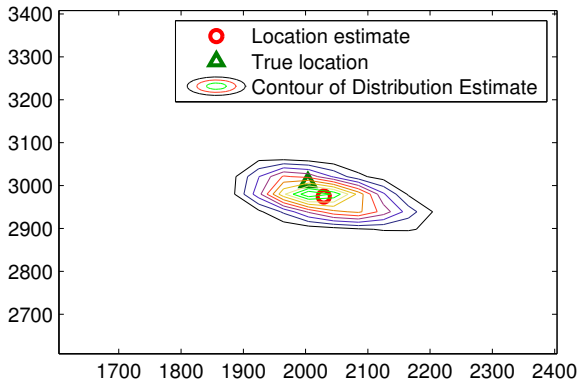

Fig. 1. RSPB distribution of location estimate, $p_{\text{nlos}} = 0$, contour plot in 10% steps

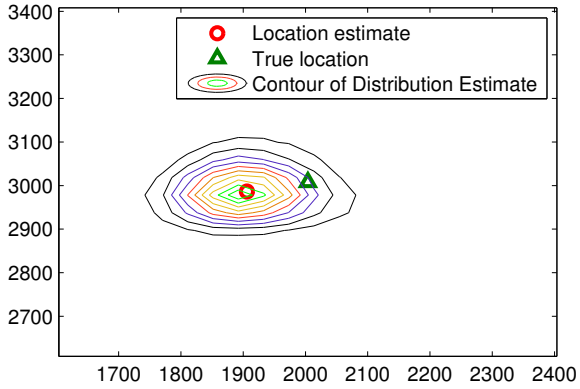

Fig. 2. RSPB distribution of location estimate, $p_{\text{nlos}} = 0.4$, contour plot in 10% steps

The distribution estimate was extracted solely through the RSPB and without additional measurements or assumptions. Due to moving MT's and changing environments, this is crucial in the context of geolocation, as a measurement cannot be repeated under the same conditions.

The distribution estimate shows, that the uncertainty in horizontal direction is larger than the uncertainty in vertical direction. The reason for this is the setup of FT's relative to the true position of the MT and the structure of outliers in the particular example.

The additional information can be utilized in a plethora of ways. In a tracking scenario, distribution estimates from past measurements can be combined to improve new position estimates. Concrete applications are subject to future research.

5. COMPARATIVE SIMULATION RESULTS

In order to assess the robustness of proposed methods and compare them to existing methods, a simulation study is con-

ducted on the regression model in Eq. (2) of order $p = 1$ with contamination $U_i \sim (1-\varepsilon)\mathcal{N}(0, \sigma_e^2)+\varepsilon\mathcal{N}(0, \kappa\sigma_e^2)$. κ was set to 10000, in order to allow for a wide range of worst case scenarios. A total of 10000 MC iterations were run over 10 levels of contamination ranging from $\varepsilon = 0$ to $\varepsilon = 0.45$, which, for the MM-estimate and a sample size of $N = 20$, is the largest amount of contamination the underlying estimator can handle. In order to reduce statistical variance, the same data set was used for all methods. Based on the resulting bootstrap distributions with $B = 100$ replicates, confidence intervals (CI's) with confidence level $\alpha = 0.9$ were obtained.

The Fast and Robust Bootstrap (FRB) is a well established robust bootstrap method ([8], [15], [13]) and is used as a benchmark, as implemented by the authors of [8].

Fig. 3 displays the Empirical Coverage Probability (ECP) as the fraction of accurate CI's, which contain β.

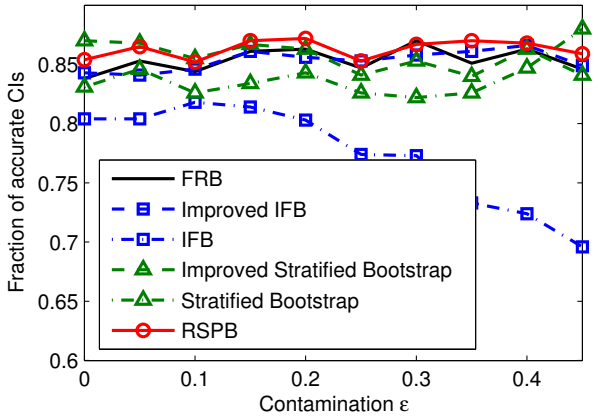

Fig. 3. Empirical Coverage Probability of discussed methods over 1000 MC iterations per contamination level

The following observations are made:

- ECP of the Improved Stratified Bootstrap is higher than the ECP of the Stratified Bootstrap, except for the case $\varepsilon = 0.45$. This is due to the early breakdown of the Stratified Bootstrap, which leads to dramatic increase in CI bias and CI length. These compensate and result in high ECP. This directly motivates the need for a second measure of robustness, discussed below.
- ECP of the Improved IFB is consistently higher than that of the IFB over all ranges of contamination, especially for large ε.
- All proposed methods perform on par with the FRB.

The breakdown point of an estimator is defined as the smallest fraction of contamination in a sample, which can cause the estimate to give information about the estimated statistic [11]. For point estimates, the maximum bias curve [16] is a well established robustness measure. A CI contains no information about the estimated statistic, if it is either infinitely biased or infinitely long. The same methods as plotted in Fig. 3 are displayed in Fig. 4 in terms of maximum CI bias and CI length. This is an extension of the classical maximum bias curve to represent the general definition of breakdown in the context of CI's.

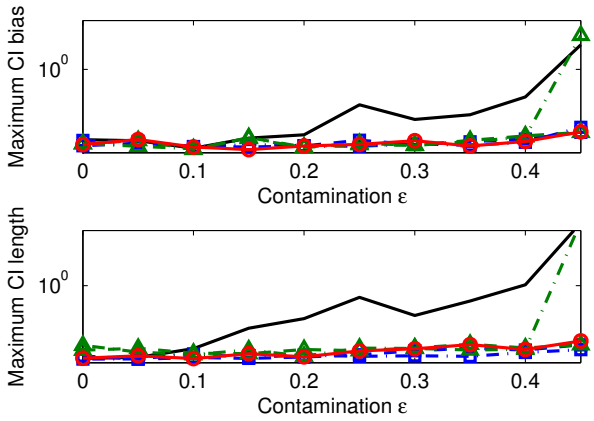

Fig. 4. Maximum CI bias and CI length over 1000 MC iterations per contamination level, same methods as in Fig. 3

The following observations are made:

- While maximum CI length and CI bias of the FRB remain bounded over all ranges of contamination, a significant increase of both statistics is present.
- The Stratified Bootstrap breaks down after $\varepsilon = 0.4$. This accounts for the misleading increase in ECP at $\varepsilon = 0.45$ in Fig. 3. The Improved Stratified Bootstrap hence outperforms the Stratified Bootstrap across all levels of contamination.
- Maximum CI bias and maximum CI length of all other methods remain small and bounded over all ranges of contamination.

6. CONCLUSION

Alterations to two existing robust bootstrap methods were proposed and their significant positive impact on the quality of so obtained CI's was demonstrated. Furthermore, a novel robust bootstrap, the Robust Starting Point Bootstrap (RSPB) was introduced and a comparative simulation study was conducted to show that it is competitive to existing methods. The RSPB was applied to the problem of regression based geolocation to extend point estimates to distribution estimates, hence extracting additional information from a given sample, without additional data or assumptions.

The project HANDiCAMS acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 323944.

7. REFERENCES

[1] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall/CRC, 1994.

[2] A. M. Zoubir and D. R. Iskander, *Bootstrap Techniques for Signal Processing*, Cambridge University Press, 2004.

[3] B. Efron, "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.

[4] A. M. Zoubir and D. R. Iskander, "Bootstrap Methods and Applications," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 10–19, 2007.

[5] A. J. Stromberg, "Robust covariance estimates based on resampling," *Journal of Statistical Planning and Inference*, vol. 57, no. 2, pp. 321–334, 1997.

[6] S. Müller and A. H. Welsh, "Outlier Robust Model Selection in Linear Regression," *Journal of the American Statistical Association*, vol. 100, no. 472, pp. 1297–1310, 2005.

[7] C. Amado and A. M. Pires, "Robust Bootstrap with Non Random Weights Based on the Influence Function," *Communications in Statistics - Simulation and Computation*, vol. 33, no. 2, pp. 377–396, 2004.

[8] M. Salibian-Barrera and R. H. Zamar, "Bootstrapping Robust Estimates of Regression," *The Annals of Statistics*, vol. 30, no. 2, pp. 556–582, 2002.

[9] U. Hammes, E. Wolsztynski, and A. M. Zoubir, "Robust Tracking and Geolocation for Wireless Networks in NLOS Environments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 5, pp. 889–901, 2009.

[10] F. Yin, C. Fritsche, F. Gustafsson, and A. M. Zoubir, "TOA-Based Robust Wireless Geolocation and Cramer-Rao Lower Bound Analysis in Harsh LOS/NLOS Environments," *IEEE Transactions on Signal Processing*, vol. 61, no. 9, pp. 2243–2255, 2013.

[11] R. A. Maronna, R. D. Martin, and V. J. Yohai, *Robust Statistics: Theory and Methods*, John Wiley & Sons, Ltd, 2006.

[12] V. J. Yohai, "High Breakdown-Point and High Efficiency Robust Estimates for Regression," *The Annals of Statistics*, vol. 15, no. 2, pp. 642–656, 1987.

[13] M. Salibian-Barrera, S. Van Aelst, and G. Willems, "Fast and Robust Bootstrap," *Statistical Methods and Applications*, vol. 17, no. 1, pp. 41–71, 2008.

[14] M. Silibian-Barrera, G. Willems, and R. Zamar, "The Fast-Tau Estimator for Regression," *Journal of Computational and Graphical Statistics*, vol. 17, no. 3, pp. 1–24, 2008.

[15] M. Salibian-Barrera and S. Van Aelst, "Robust model selection using fast and robust bootstrap," *Computational Statistics & Data Analysis*, vol. 52, no. 12, pp. 5121–5135, 2008.

[16] J. R. Berrendero, B. V. M. Mendes, and D. E. Tyler, "On the Maximum Bias Functions of MM-Estimates and Constrained M-Estimates of Regression," *The Annals of Statistics*, vol. 35, no. 1, pp. 13–40, 2007.