SOURCE NUMBER ESTIMATION IN IMPULSIVE NOISE ENVIRONMENTS USING
BOOTSTRAP TECHNIQUES AND ROBUST STATISTICS

Zhihua Lu, Yacine Chakhchoukh and Abdelhak M. Zoubir

Signal Processing Group
Technische Universität Darmstadt
Merckstrasse 25, Darmstadt 64283, Germany.
Email: zlu, ychakh, zoubir@spg.tu-darmstadt.de

ABSTRACT
We consider the problem of source number estimation in array processing when impulsive noise is present. To combat impulsive noise more effectively, two robust estimators with high breakdown points, i.e., the minimum covariance determinant (MCD) estimator and the MM-estimator are applied in combination with the bootstrap. The MCD estimator is applied to discard the outliers in observations caused by impulsive noise, while the MM-estimator is suggested to estimate robustly the covariance matrix of observations. Simulations show the significant performance gain offered by the two proposed methods in terms of correctly estimating the number of sources at low SNR in the presence of impulsive noise.

Index Terms— array processing, source number estimation, impulsive noise, bootstrap, robust statistics.

1. INTRODUCTION

Estimation of the number of sources impinging on an array of sensors is one of the most fundamental problems in array processing. The classical methods are based on information theoretic criteria (ITC), including Akaike’s information criterion (AIC) and Rissanen’s minimum description length criterion (MDL) [2]. They rely on strict distributional assumptions, e.g., Gaussian sources and Gaussian noise. However, impulsive noise has been considered as a more accurate description for ambient noise in many communication channels such as urban and indoor radio channels, due to the impulsive nature of man-made electromagnetic interference and a large amount of natural noise [3]. The presence of impulsiveness causes often large performance degradation of ITC based methods. Besides, they also do not give satisfactory performance in some severe practical environments such as small sample size and low SNR regime.

The bootstrap-based multiple hypotheses test proposed in [1] relaxes the assumption of Gaussian data and is, to some extent, more robust than ITC based methods. However, its performance is very poor in low SNR regimes with the presence of impulsive noise causing outliers in real data. This is because, like most of the existing methods, the eigenvalues are obtained from the sample covariance (e.g., Eq. (4)) of observations. The sensitivity of the sample covariance to outliers leads to the inaccuracy of eigenvalues estimation. More importantly, the bootstrap has a very low breakdown point since the bootstrap distribution may be severely affected by bootstrap samples with a higher proportion of outliers than the one in the original data set. To improve the robustness of the covariance estimator and prevent expanding outlier contamination in bootstrap samples, two robust estimators with a high breakdown point are adopted and combined with bootstrap techniques. Thus, two new methods are proposed to estimate the number of sources in the context of small sample size, low SNR and heavy impulsive noise. In the first proposed method, the minimum covariance determinant (MCD) estimator [4] is used to detect and discard outliers caused by impulsive noise. Then the outlier-free data are processed by the bootstrap-based method in [1]. In the second proposed method, the traditional bootstrap is replaced by the fast and robust bootstrap (FRB)[7], where the MM-estimator is integrated.

The remainder of the paper is organized as follows. The array signal model is introduced briefly in Section 2, followed by a description of the bootstrap based method for source number estimation in Section 3. Two robust methods are introduced and discussed in Section 4. Simulation results are given in Section 5, before conclusions are drawn in Section 6.

2. ARRAY SIGNAL MODEL

Consider q narrowband far-field sources impinging on an array with p sensors (p > q). The received n snapshots of independent and identically distributed (i.i.d) circular complex data could be written as

\[ x_i = As_i + n_i, \quad i = 1, \ldots, n \]  

(1)

The work was supported by the Graduate School of Computer Engineering at Technische Universität Darmstadt and Deutsche Forschungsgemeinschaft (DFG).
where $A$ is the $p \times q$ array steering matrix, $s_i$ is the $q$-dimensional source signal with zero mean, and $n_i$ is the source-independent i.i.d. noise with zero mean and covariance $\sigma_i^2 I$. The covariance of received data is given by

$$R_e = E[x_ix_i^H] = AR_e A^H + \sigma_i^2 I$$  \hspace{1cm} (2)

where $R_e = E[s_is_i^H]$ is the source covariance, and $(\cdot)^H$ is the Hermitian transpose. The eigenvalues of $R_e$ are given by

$$\lambda_1 \geq \cdots \geq \lambda_q \geq \lambda_{q+1} = \cdots = \lambda_p = \sigma_i^2$$  \hspace{1cm} (3)

where the first $q$ eigenvalues belong to the source signal, and the last $p - q$ to the noise. In reality, only a finite number of snapshots is available to calculate the sample covariance matrix, which is given by

$$\hat{R}_e = \frac{1}{n} \sum_{i=1}^{n} x_ix_i^H$$  \hspace{1cm} (4)

with corresponding eigenvalues:

$$\hat{\lambda}_1 > \cdots > \hat{\lambda}_q > \hat{\lambda}_{q+1} > \cdots > \hat{\lambda}_p.$$  \hspace{1cm} (5)

However, the differences between noise eigenvalues are relatively smaller than those between source eigenvalues. In general, the problem of source number estimation is addressed based on counting the multiplicity of the smallest eigenvalues. In the sequel, we focus on the bootstrap-based method proposed in [1] due to its superiority over ITC based methods for the small sample size case.

3. THE BOOTSTRAP-BASED MULTIPLE HYPOTHESES TEST

In order to determine the number of the smallest eigenvalues that are equal, the following set of hypotheses is constructed:

$$H_0 : \lambda_1 = \cdots = \lambda_p$$
$$\vdots$$
$$H_k : \lambda_{k+1} = \cdots = \lambda_p$$
$$\vdots$$
$$H_{p-2} : \lambda_{p-1} = \lambda_p$$  \hspace{1cm} (6)

with corresponding alternatives $K_0, K_k, \ldots, K_{p-2}$. The total test procedure in Table 1 is given in [1]. Each hypothesis is obtained by

$$H_k = \bigcap_{i,j} H_{ij}, \quad i = k + 1, \ldots, p - 1, \quad j = i + 1, \ldots, p$$  \hspace{1cm} (7)

where $H_{ij} : \lambda_i = \lambda_j$ tests the difference of two eigenvalues. The test result of $H_k$ is derived based on Bonferroni’s multiple tests procedure which tests the group of hypotheses $H_{ij}$’s simultaneously. The null distributions of the test statistics are estimated by employing the bootstrap resampling algorithm [9].

<table>
<thead>
<tr>
<th>Step 1</th>
<th>Set $k = 0$.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 2</td>
<td>Test $H_k$.</td>
</tr>
<tr>
<td>Step 3</td>
<td>If $H_k$ is accepted then set $\hat{k} = k$ and stop.</td>
</tr>
<tr>
<td>Step 4</td>
<td>If $H_k$ is rejected and $k &lt; p - 1$ then set $k = k + 1$ and return to Step 2. Otherwise set $\hat{k} = p - 1$ and stop.</td>
</tr>
</tbody>
</table>

4. THE ROBUST BOOTSTRAP-BASED METHODS

The ambient noise $n_i$ in Eq. (1) is generally impulsive, which can be modeled by the $\varepsilon$-contaminated Gaussian mixture model with probability density function (pdf)

$$f = (1 - \varepsilon)N(0, \nu^2) + \varepsilon N(0, \kappa \nu^2).$$  \hspace{1cm} (8)

where $\nu > 0$, $0.01 < \varepsilon \leq 0.1$ and $10 < \kappa < 100$. Here, $\varepsilon$ denotes the probability that impulses occur. The deviation from the Gaussian assumption of noise has a large distorting influence on the covariance estimator in Eq. (4), causing biased estimation of the number of sources. The approach consists in using robust statistics that are resistant against impulsive noise effects. In array processing, the received observation $x_i$ is complex valued. Herein, robust statistics are applied to the real vector $z_i = [\text{Re}(x_i^T), \text{Im}(x_i^T)]^T$, which is a concatenation of the real and imaginary parts of the complex vector $x_i$. Then, the covariance estimator of $x_i$ can be recovered from that of $z_i$.

For the first proposed method, we apply a procedure of outlier detection as a pre-processing step, for which the robust Mahalanobis distance is calculated by involving the minimum covariance determinant (MCD) estimator [4]. The MCD estimator is based on $h$ observations whose classical covariance matrix has the lowest determinant. A standard choice for $h$ is $[(n + p + 1)/2]$ which yields the maximal breakdown point$^1$. Herein, these $h$ observations are found by a fast algorithm called FAST-MCD [5]. The MCD estimate of location is then the average of these $h$ points, and the MCD estimate of scatter is their covariance matrix. The main steps of the first proposed method are given in Table 2.

For the second proposed method, we use an MM-estimator, which is both highly robust against outliers and highly efficient for normal data [8]. It can be expressed as the solution of a fixed point equation [6]:

$$\hat{\Theta}_n = f_n(\hat{\Theta}_n)$$  \hspace{1cm} (9)

where $\hat{\Theta}_n$ is a vector containing the MM-estimator of interest, e.g., location and shape. If the MM-estimator is used to calculate the covariance matrix of observations instead of the sample covariance estimator in Eq. (4), the result will be

$^1$The breakdown point of an estimator is the proportion of outliers an estimator can handle before giving an arbitrarily large result.
more robust against outliers. However, simply re-calculating MM-estimators for all bootstrap samples, i.e.,

$$\hat{\Theta}_n^* = f_n^*(\hat{\Theta}_n^*)$$

(10)

will cause two main problems. Firstly, high computational cost would be needed for solving Eq. (10) for the required number of bootstrap samples. Secondly, in some bootstrap samples, the proportion of outliers which is much higher than the original data set, exceeds the breakdown value of the MM-estimator. And thus, it leads to unreliable MM-estimators for such bootstrap samples. To integrate the MM-estimator with the bootstrap compactly, the references [6] [7] proposed an algorithm called the fast and robust bootstrap (FRB). The MM-estimator for the bootstrap sample is given by

$$\hat{\Theta}_n^* := \hat{\Theta}_n + [I - \nabla f_n(\hat{\Theta}_n)]^{-1}(f_n^*(\hat{\Theta}_n) - \hat{\Theta}_n)$$

(11)

where the MM-estimator $\Theta_n^*$ for bootstrap samples is approximated by the MM-estimator $\hat{\Theta}_n$ of the original observations which is calculated in Eq. (9). Since the bootstrap sample of observations is only used to compute the function $f_n^*$. Based on a linear correction in Eq. (11), computation of Eq. (10) is avoided. The main steps of the second proposed method are given in Table 3.

### Table 2. The method based on the MCD estimator and the bootstrap.

| Step 1: | Construct the real observations $Z = \{z_1, \ldots, z_n\}$ from the complex observations $X = \{x_1, \ldots, x_n\}$. |
| Step 2: | Obtain the sample mean $T_n$ and the covariance matrix $C_n$ based on $h$ real observations derived by the FAST-MCD method. |
| Step 3: | Calculate the robust Mahalanobis distances for $z_i$: $\text{RD}(z_i) = \sqrt{(z_i - T_n)^T C_n^{-1} (z_i - T_n)}$, $i = 1, \ldots, n$. |
| Step 4: | Compare the robust Mahalanobis distances with a cutoff value $\sqrt{\chi^2_{2p, 0.975}}$, and discard all the $z_i$'s (and corresponding $x_i$'s) with higher distance values than the cutoff value which are considered as outliers. |
| Step 5: | Get the eigenvalues of the sample covariance (in Eq. (4)) of outlier-free complex-valued observations. |
| Step 6: | Resample outlier-free complex observations, get the bootstrap samples of eigenvalues and construct the null distribution of the test statistics. |
| Step 7: | Run the sequential test procedure in Table 1. |

### Table 3. The method based on the fast and robust bootstrap.

| Step 1: | Construct the real observations $Z = \{z_1, \ldots, z_n\}$ from the complex observations $X = \{x_1, \ldots, x_n\}$. |
| Step 2: | Calculate the MM-estimator of $Z$ using Eq. (9), from which recover the MM-estimator of $X$. |
| Step 3: | Get the eigenvalues from the MM-estimator of $X$. |
| Step 4: | Resample $Z$, calculate the MM-estimators $\hat{\Theta}_n^*$ for the bootstrap samples $Z^*$ using Eq. (11), from which recover the MM-estimators $\Theta_n^*$ for the bootstrap samples $X^*$. |
| Step 5: | Get the bootstrap samples of eigenvalues from $\Theta_n^*$ and construct the null distribution of the test statistics. |
| Step 6: | Run the sequential test procedure in Table 1. |

### 5. SIMULATIONS

An uniform linear array with inter-sensors spacing of half the wavelength was employed. Simulation results were obtained based on 100 snapshots of a Gaussian source signal contaminated by impulsive noise, and 500 Monte Carlo runs. The number of bootstrap samples was chosen as $B = 200$, and a global level of significance $\alpha = 2\%$ was set for the multiple hypotheses test. The SNR range in the experiment was focused on $[-16, 0]$ dB and $\kappa = 100$ was set for high impulsiveness. Since the percentage of impulsive noise is assumed to be less than 10\% in our scenario, we choose $h = 0.75n$ for the MCD estimator to gain some efficiency. The breakdown value of 0.5 and the efficiency of 95\% were set for the MM-estimator. The traditional bootstrap based method [1] is denoted by “BTS”. The two robust methods based on the MCD estimator and the FRB procedure are denoted by “MCD” and “FRB”, respectively.

Suppose that we have an array with 4 sensors and 2 sources, which are located at 20° and 40° with respect to broadside. The results are quantified by the empirical probability of correctly detecting the source number vs. SNR, see Figs. 1, 2 and 3. In Fig. 1, $\varepsilon = 0$ means the noise is Gaussian. Both the MCD and the FRB show large performance gain at very low SNR compared to the BTS method, although small performance loss is observed at moderate SNR. In Fig. 2, $\varepsilon = 1\%$ means a small fraction of impulsiveness. Both proposed methods offer significant performance gain over the BTS throughout all the SNR regime being tested, especially for the FRB at very low SNR. Moreover, The BTS suffers large performance loss compared to the case of Gaussian noise while the two proposed methods preserve their performance quite well. In Fig. 3, $\varepsilon = 10\%$ indicates a large fraction of impulsiveness. In this case, although both proposed methods still outperform the BTS which breaks down totally,
it is apparent that their performance decreases compared to the case of Gaussian noise or small fraction of impulsiveness. Generally speaking, with impulsive noise getting heavier, the performance of the BTS degrades vary fast; the FRB preserves its performance better than the MCD at very low SNR (especially for a small fraction of impulsiveness), and the inverse is true at a moderate SNR.

![Fig. 1. Detection rate in Gaussian noise.](image1)

![Fig. 2. Detection rate for a small fraction of impulsiveness.](image2)

### 6. CONCLUSION

We have addressed the problem of source number estimation in some severe practical situations of interest, i.e., small sample size, low SNR and heavy impulsive noise. More precisely, two robust estimators, which are highly resistant to outliers, i.e., the MCD estimator and the MM-estimator, have been incorporated into a bootstrap-based method. The two proposed methods are very effective to suppress impulsive noise effects at low SNR, since they inherit both the advantage of the bootstrap, i.e., distributional assumption relaxation and that of robust statistics, i.e., resistance towards outliers. The computational cost for the proposed methods is only slightly increased due to the fact that robust estimators are computed once for the original data rather than for all the bootstrap samples.

### 7. REFERENCES