SOURCE ENUMERATION USING THE BOOTSTRAP FOR VERY FEW SAMPLES

Zhihua Lu and Abdelhak M. Zoubir

Signal Processing Group
Technische Universität Darmstadt
Merckstrasse 25, Darmstadt 64283, Germany
phone: + 49 6151-16 4595, fax: + 49 6151-16 3778
email: {zlu, zoubir}@spg.tu-darmstadt.de
web: www.nt.tu-darmstadt.de/spg

Florian Roemer and Martin Haardt

Communications Research Laboratory
Ilmenau University of Technology
P. O. Box 100565, D-98694 Ilmenau, Germany
phone: +49 (3677) 69-2613, fax: +49 (3677) 69-1195
email: {florian.roemer, martin.haardt}@tu-ilmenau.de
web: www.tu-ilmenau.de/crl

ABSTRACT

We consider the problem of source enumeration in array processing when only few samples are available. In this case, the noise eigenvalues[1] spread, so that most existing methods, which assume equality of the noise eigenvalues implicitly, suffer large performance loss or even break down. We present a method based on hypothesis testing with the bootstrap. The test statistic is derived by using the exponential profile property of the noise eigenvalues. Simulations show the significant performance gain offered by the proposed method in terms of correctly detecting the number of sources for a very small sample size.

1. INTRODUCTION

Inferring the number of sources impinging on an array of sensors is the critical first step in a subsequent signal parameter estimation in array processing. Because of computational and modeling simplicity, most classical methods are derived based on the sample eigenvalues of the sample covariance matrix. One category of classical methods is based on information theoretic criteria (ITC), including Akaike's information criterion (AIC) and Rissanen's minimum description length criterion (MDL), proposed by Wax and Kailath [1]. The other category is based on sequential hypothesis testing procedures, including the sphericity test [2] and the bootstrap-based test [3]. Addressing the problem of source enumeration based on the asymptotic distributions of the sample eigenvalues provided by random matrix theory has emerged recently [5]–[7].

Of interest is to detect the number of sources using a number of samples comparable to or even smaller than the system size, i.e., the number of sensors, which has become increasingly important in many state-of-the-art radar and sonar systems. When the sample size is extremely small, the spread of the noise eigenvalues is quite significant [5]. Thus, the noise eigenvalues are not sufficiently close to each other. A systematic description of the spreading phenomenon is given in [8], which is known as the Marčenko-Pastur density. Unfortunately, most existing methods either use some asymptotic distributions of the sample eigenvalues, which are inaccurate for relatively few samples, or ignore the spreading phenomenon of the noise eigenvalues, so that they do not yield satisfactory performance in such severe practical situation. New methods which take into account the blurring of the noise eigenvalues are expected to emerge.

More recently, the authors in [5] developed a new method for relatively few samples, by using the Marčenko-Pastur density. This method shows the superiority with respect to the Wax-Kailath MDL-based method [1]. The assumptions, such as high-dimensional spiked signal, asymptotic regime for the system size and the sample size, render the method inappropriate for all cases of small sample size.

In [3], bootstrap techniques are adopted to construct a sequential hypothesis testing procedure, for testing the equality of the sample eigenvalues. It performs well when the sample size is not extremely small, but its computational complexity is exponentially proportional to the system size. It is inapplicable for the high-dimensional array system. Besides, the test statistic is inappropriate, since equality of the noise eigenvalues does not hold for very few samples due to the spreading of the noise eigenvalues. To address these two issues, we use a more accurate test statistic which reflects fluctuations of the noise eigenvalues, following the heuristic result in [4], that is, the profile of the ordered noise eigenvalues is seen to approximately fit an exponential law for white Gaussian noise and short data. Then, a relatively computationally simple bootstrap-based test procedure is constructed in order to infer the number of sources.

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 in [3] in Section 3. The main idea of the proposed method is given in Section 4. Simulation results and short discussions are given in Section 5, before conclusions are drawn in Section 6.

2. ARRAY SIGNAL MODEL

Consider q narrow-band 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 can be written as

$$\mathrm{x}_i = \boldsymbol{A}\boldsymbol{s}_i + \boldsymbol{v}_i,\ i = 1,\ldots,n \tag{1}$$

where $\boldsymbol{A}$ is the $p \times q$ array steering matrix, $\boldsymbol{s}_i$ is the q-dimensional source signal with zero mean, and $\boldsymbol{v}_i$ is the source-independent i.i.d. noise with zero mean and covariance $\sigma_v^2\boldsymbol{I}$. The population covariance matrix of the received data is given by

$$\boldsymbol{R}_x = \mathrm{E}[\mathrm{x}_i\mathrm{x}_i^H] = \boldsymbol{A}\boldsymbol{R}_s\boldsymbol{A}^H + \sigma_v^2\boldsymbol{I} \tag{2}$$

where $\boldsymbol{R}_s = \mathrm{E}[\mathrm{s}_i\mathrm{s}_i^H]$ is the source covariance. $(\cdot)^H$ and $\boldsymbol{I}$ denote the Hermitian transpose and the identity matrix, re-

The work was supported by Deutsche Forschungsgemeinschaft (DFG) CAMUS project and the Graduate School of Computational Engineering at Technische Universität Darmstadt.

[1] Without special statement, the eigenvalues mentioned in the sequel denote the sample eigenvalues, calculated from a sample covariance matrix.

spectively. The population eigenvalues of $\boldsymbol{R}_x$ are given by

$$\lambda_1 \geq \cdots \geq \lambda_q \geq \lambda_{q+1} = \cdots = \lambda_p = \sigma_v^2 \tag{3}$$

where the first q eigenvalues belong to the source signal, and the last $p-q$ to the noise. In general, the problem of source enumeration is addressed based on counting the multiplicity of the smallest eigenvalues. However, only a finite number of snapshots is available in reality, so that we do not have access to the population covariance matrix but to its finite sample estimate

$$\hat{\boldsymbol{R}}_x = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^H \tag{4}$$

with corresponding sample eigenvalues:

$$l_1 > \cdots > l_q > l_{q+1} > \cdots > l_p \tag{5}$$

Which are all distinct with probability one. Although the joint distributions of the sample eigenvalues are given in some different forms in the case of Gaussian data, they are mathematically intractable or unreliable for the small sample size case. In order to avoid using these cumbersome distributions for the problem of source enumeration, the bootstrap is proposed to simulate the distribution of the sample eigenvalues, more precisely, the distribution of the sufficient test statistic constructed based on the sample eigenvalues. In the next section, we focus on the bootstrap-based method proposed in [3] due to its superiority over ITC-based methods for the small sample size case.

3. THE BOOTSTRAP-BASED TEST

Assuming that the differences between the noise sample eigenvalues are relatively smaller than those between the source sample eigenvalues, in order to detect statistically significant difference of the eigenvalues, the following set of hypotheses is constructed:

$$\begin{array}{llll} \mathsf{H}_0 & : \lambda_1 & = & \cdots = \lambda_p \\ \vdots & & \vdots & \\ \mathsf{H}_k & : \lambda_{k+1} & = & \cdots = \lambda_p \\ \vdots & & \vdots & \\ \mathsf{H}_{p-2} & : \lambda_{p-1} & = & \lambda_p \end{array} \tag{6}$$

with corresponding alternatives $\mathsf{K}_0, \mathsf{K}_k, \ldots, \mathsf{K}_{p-2}$. Each hypothesis is obtained by

$$\mathsf{H}_k = \bigcap_{i,j} \mathsf{H}_{ij}, \quad i = k+1, \ldots, p-1,\ j = i+1, \ldots, p \tag{7}$$

where $\mathsf{H}_{ij} : \lambda_i = \lambda_j$ tests the difference of two eigenvalues. These hypotheses are tested sequentially, starting from H_0 until finding a true hypothesis. If the hypothesis H_k is rejected, move the test forward to the next hypothesis H_{k+1}. Otherwise, accept the hypothesis H_k and stop the total test procedure. As the test result, acceptance of H_k indicates the estimate $\hat{q} = k$.

The hypotheses in Eq. (6), which are constructed by intersections between sub-hypotheses in Eq. (7), are tested with a multiple hypotheses test (MHT) [11]. It means that 2^p hypotheses are tested simultaneously no matter how many sources are present. The computational cost increases exponentially with the array size p. Also, the bootstrap resampling algorithm [10] which is a computer-intensive method is employed to simulate the null distribution of the test statistic for each hypothesis. Bootstrap-based MHT formulations can be computationally expensive, especially for the high-dimensional array system. Therefore, a computationally simple test procedure is to be sought for.

This method which tests the equality of the sample eigenvalues, performs efficiently when the number of samples n is relatively large, e.g., $n = 100$. When n is comparable to p, e.g., $n = p = 10$, the method breaks down due to the noise eigenvalues spread. To remedy this problem, [3] introduced the concept of bias of the sample eigenvalues, following the result of [12]. It is necessary to reduce the bias which becomes quite significant in the very small sample size case. Through bias correction for the sample eigenvalues, the assumption that the noise-only sample eigenvalues have equal means, to some extent, is recalled. The bias-corrected bootstrap-based test continued to work for very few samples.

4. THE MODIFIED BOOTSTRAP-BASED TEST

In this section, instead of using bias correction before the tests, we use a more appropriate test statistic which considers the fluctuations of the noise eigenvalues due to the very small sample size. The authors in [4] show the approximate exponential profile of the ordered noise eigenvalues (see Eq. (5))

$$l_\alpha = l_\beta r_{m,n}^{\alpha-\beta}, \quad \alpha, \beta = q+1, \ldots, p \tag{8}$$

where $r_{m,n} = e^{-2a} (a > 0)$ denotes the exponential function of the number of the noise eigenvalues $m = p - q$ and the number of samples n. Thus the sequence of the ordered noise eigenvalues seems to be a geometric series. a can be derived based on the assumption,

$$\sum_{i=1}^{m} l_i = m\sigma_v^2 \tag{9}$$

where σ_v^2 denotes the noise variance, and an order-4 Taylor expansion of the hyperbolic tangent function. A corrected version [9] which removes the assumption $m \leq n$ is given as

$$a = \sqrt{\frac{1}{2}\left(\frac{15}{\mu^2+2} - \sqrt{\frac{225}{(\mu^2+2)^2} - \frac{180\mu}{\nu(\mu^2-1)(\mu^2+2)}}\right)} \tag{10}$$

where $\mu = \min\{m, n\}$ and $\nu = \max\{m, n\}$. It can be seen from the computation of $r_{m,n}$ that the relationship in Eq. (8) is valid for all sample sizes, with the extreme case of the noise eigenvalues becoming equal as n tends to infinity. Due to the relationships in Eqs. (8) and (9), from preceding smaller observed noise eigenvalues, we can predict the next noise eigenvalue:

$$\tilde{l}_{p-i} = (i+1)\frac{1 - r_{i+1,n}}{1 - (r_{i+1,n})^{i+1}} \hat{\sigma}_v^2, \quad i = 1, \ldots, p-1 \tag{11}$$

with

$$\hat{\sigma}_v^2 = \frac{1}{i}\sum_{j=0}^{i-1} l_{p-j} \tag{12}$$

where $\hat{\sigma}_v^2$ is an estimator of the noise variance, according to Eq. (9). Then, we test the hypothesis

$$\begin{aligned} \mathsf{H}_i &: \lambda_{p-i} = \tilde{l}_{p-i} \quad \text{against} \\ \mathsf{K}_i &: \lambda_{p-i} \neq \tilde{l}_{p-i} \end{aligned} \tag{13}$$

where λ_{p-i} is the population eigenvalue with its estimate l_{p-i} which is obtained from the sample covariance matrix. The test is conducted by the bootstrap, see Table 1 [10], where $\theta = \lambda_{p-i}$, $\hat{\theta} = l_{p-i}$ and $\theta_0 = \tilde{l}_{p-i}$. If H_i is accepted, the observed noise eigenvalue still follows the theoretical exponential profile, that is, λ_{p-i} belongs to one of the noise eigenvalues. Otherwise, λ_{p-i} is one of the source eigenvalues. Following this statement, we construct a sequential test procedure in Table 2, in order to detect the number of the noise or source eigenvalues.

Table 2. The sequential test procedure.

Step 1. Set $i = 1$.
Step 2. Test the hypothesis in Eq. (13).
Step 3. If K_i is accepted then set $\hat{q} = p - i$ and stop.
Step 4. If H_i is accepted and $i < p - 1$ then set $i = i + 1$ and return to Step 2. Otherwise set $\hat{q} = 0$ and stop.

Table 1. The bootstrap-based test for the hypothesis $\mathsf{H} : \theta = \theta_0$ against $\mathsf{K} : \theta \neq \theta_0$.

Step 0. *Experiment.* Conduct the experiment and collect the data into the sample $\mathscr{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. Calculate the test statistic

$$T_n = |\hat{\theta} - \theta_0| / \hat{\sigma},$$

where $\hat{\theta}$ is an estimator of θ and $\hat{\sigma}^2$ is an estimator of the variance σ^2 of $\hat{\theta}$.
Step 1. *Resampling.* Draw a random sample of size n, with replacement from $\mathscr{X}$

$$\mathscr{X}^* = \{\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_n^*\}.$$

Step 2. *Calculation of the bootstrap statistic.* From $\mathscr{X}^*$, calculate

$$T_n^* = |\hat{\theta}^* - \hat{\theta}| / \hat{\sigma}^*,$$

where $\hat{\theta}^*$ and $\hat{\sigma}^*$ are computed in the same manner as $\hat{\theta}$ and $\hat{\sigma}$, but with the bootstrap sample $\mathscr{X}^*$ replacing $\mathscr{X}$.
Step 3. *Repetition.* Repeat Steps 1 and 2 many times to obtain a total of B bootstrap estimates $T_{n,1}^*, T_{n,2}^*, \cdots, T_{n,B}^*$.
Step 4. *Ranking.* Rank the collection $T_{n,1}^*, T_{n,2}^*, \ldots, T_{n,B}^*$ into increasing order to obtain

$$T_{n,(1)}^* \leq T_{n,(2)}^* \leq \cdots \leq T_{n,(B)}^*.$$

Step 5. *Test.* A bootstrap test has then the following form: reject H if $T_n > T_{(q)}^*$, where the choice of q determines the level of significance of the test and is given by $\alpha = (B + 1 - q)(B + 1)^{-1}$, where α is the nominal level of significance.

It is worth mentioning that the exponential fitting test (EFT) proposed in [4] used a more complicated test statistic, whose distribution is unknown. For this reason, the threshold for the hypothesis test was calculated by Monte Carlo simulations with a prior knowledge of the exact noise distribution. It is unrealistic in practice since it is not always possible to repeat the experiment for data collection or there is not enough a prior knowledge to run Monte Carlo simulations. In this case, the bootstrap is a proper alternative, due to its simple and attractive properties. This is validated by the simulations in Section 5.

5. SIMULATIONS AND DISCUSSIONS

A uniform linear array with inter-sensors spacing of half the wavelength was employed. For simplicity, the case of uncorrelated Gaussian source signals contaminated by Gaussian noise was considered. Simulation results were obtained based on 500 Monte Carlo runs. The number of bootstrap samples was chosen as $B = 200$, and a level of significance $\alpha = 2\%$ was set for all involved hypothesis tests. The traditional bootstrap-based methods [3] without and with bias correction are denoted by “BTSeqn” and “BTSbas”[2], respectively. Denote the methods proposed in [4] and [5] by “EFT” and “NAD”, respectively. The method proposed in this paper is denoted by “BTSexp”.

Suppose that we have an array with 8 sensors and 3 sources, which are located at $-10°, 5°, 15°$ with respect to broadside. The signal-to-noise ratio (SNR) range in this simulation was focused on $[0, 16]$ dB. Only 10 snapshots were used. The results are quantified by the empirical probability of correctly detecting the source number (i.e., Detection rate) vs. SNR, see Fig. 1. For very few samples, the method BTSeqn breaks down completely. With bias correction of the eigenvalues, the method BTSbas starts to work, although it performs unsatisfactorily. The method EFT, with knowing the exact noise distribution a priori, has slightly lower detection rate than the well known method NAD. The proposed method BTSexp performs better than the other methods. It has highest convergence rate with respect to SNR. It is worth noting that the method BTSexp suffers a large performance degradation at low SNRs (e.g., SNR < 0 dB).

Suppose that we have an array with 15 sensors and 3 sources, which are located as in the preceding simulation. The SNR was set as 6 dB for all sources. The number of snapshots n varied in $[10, 18]$. The detection rate is given with respect to the number of snapshots n in Fig. 2. In this setting, the method BTSexp outperforms significantly the other two methods. Its performance is quite stable as the number of samples n increases. Unlike the method NAD, the method EFT decreases its performance as n increases. It seems that Monte Carlo simulations fail to provide accurate knowledge of the distribution of the test statistic, especially when n is relatively large.

Based on the above simulations for the very few sample case, we can see that the proposed method BTSexp provides the best results in terms of source enumeration at a relatively high SNR when the number of samples are close

[2] Herein, the jackknife is used for bias correction. More details about the jackknife are introduced in [14].

to the number of sensors. It increases the performance gain substantially compared to the method BTSbas or BTSeqn, while reducing the computational cost. The involved new test statistic is more efficient in dealing with the spreading phenomenon of the noise eigenvalues than bias correction for the sample eigenvalues. It is also apparent that the bootstrap is a much better choice than Monte Carlo simulations for inferring the statistics numerically in our case, since the method BTSexp is superior to the method EFT. In addition, compared to the method NAD which has the lowest computational complexity, the minimal distributional assumptions are made for the method BTSexp.

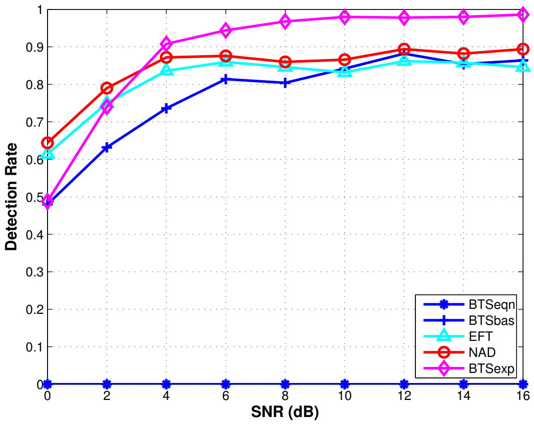

Fig. 1. Detection rate vs. SNR.

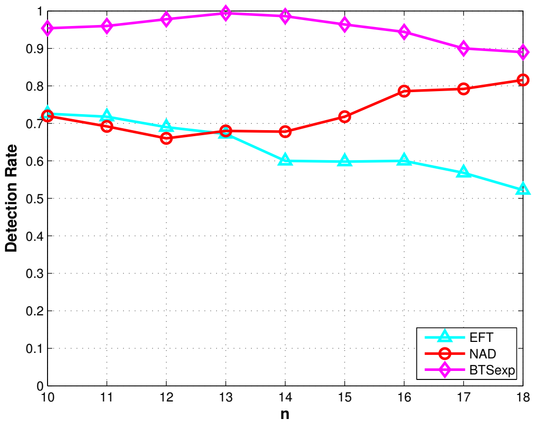

Fig. 2. Detection rate vs. number of samples.

6. CONCLUSIONS

The problem of source enumeration was investigated from a hypothesis testing viewpoint for the case of very few samples, i.e., the sample size is nearly equal to the array size. In light of the spreading phenomenon of the noise eigenvalues for very few samples, the property of the noise eigenvalues' exponential profile was used to construct the test statistic. Then, the null distribution of the test statistic was provided via bootstrap techniques, avoiding the use of cumbersome distributions of the sample eigenvalues. Finally, the number of sources was detected through a sequence of hypothesis tests. Simulations show that the proposed method outperforms the original bootstrap-based method and the well known method proposed in [5]. For future work, a more efficient estimator of noise variance $\hat{\sigma}_v^2$ (e.g., [6]) is expected to replace the one in Eq. (12), which is originally designed for the infinite sample size [13]. More importantly, a more accurate test statistic is to be found, with the help of random matrix theory.

REFERENCES

[1] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 387-392, Apr. 1985.

[2] D.B. Williams and D.H. Johnson, "Using the sphericity test for source detection with narrow-band passive arrays," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 9, pp. 2008-2014, Nov. 1990.

[3] R.F. Brcich, A.M. Zoubir and P. Pelin, "Detection of sources using bootstrap techniques," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. .206-215, Feb. 2002.

[4] A. Quinlan, J.P. Barbot, P. Larzabal and M. Haardt, "Model order selection for short data: An exponential fitting test," *EURASIP J. Adv. Signal Process.*, vol. 2007, Article ID 71953.

[5] R. Nadakuditi and A. Edelman, "Sample eigenvalues based detection of high-dimensional signals in white noise using relatively few samples," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 2625-2638, Jul. 2008.

[6] S. Kritchman and Boaz Nadler, "Non-parametric detection of the number of signals: hypotheses testing and random matrix theory," *IEEE Trans. Signal Process.*, vol. 57, no. 10, pp. 3930-3941, Oct. 2009.

[7] R. Couillet, J. W. Silverstein, Z. Bai and M. Debbah, "Eigen-Inference for Energy Estimation of Multiple Sources," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2420-2439, 2011.

[8] V.A. Marcĕnko and L.A. Pastur, "Distribution of eigenvalues in certain sets of random matrices," *Mat. Sb. (N.S.)*, vol. 72, no. 114, pp. 507-536, 1967.

[9] J. P. C. L. da Costa, M. Haardt, F. Rmer, and G. Del Galdo, "Enhanced model order estimation using higher-order arrays," in *Proc. 41-st Asilomar Conf. on Signals, Systems, and Computers*, pp. 412-416, Nov. 2007, invited paper.

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

[11] Y. Hochberg and A. Tamhane, *Multiple Comparison Procedures*, New York: Wiley, 1987.

[12] D. Lawley, "Tests of significance for the latent roots of covariance and correlation matrices," *Biometrika*, vol. 43, pp. 128-136, 1956.

[13] T.W. Anderson, "Asymptotic theory for principal component analysis," *Ann. J. Math. Stat.*, vol. 34, pp. 122-148, 1963.

[14] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, 1993.