A method for source number estimation with density-based clustering algorithm

LI Qi1, NING Jinzhi1 and DOU Jiao1
1DFH Satellite Co. Ltd., Beijing, China
E-mail: liqi2008bit@163.com

Abstract. Almost all high-performance spatial spectrum estimation algorithms for direction of arrival (DOA) estimation are based on accurate source number estimation. Applied with features of covariance matrix eigenvalues of received data and compared with commonly-used K-means algorithm, this article proposes to use a density-based clustering algorithm, which is ordering points to identify the clustering structure (OPTICS) and its detailed steps, to estimate the accurate source number. The experimental results applying with OPTICS algorithm demonstrate better performance on source number estimation over other algorithms under the same conditions.

1. Introduction

The accuracy of source number estimation is crucial to the orthogonal calculation of signal and noise subspace, when the spatial spectrum estimation method is used for direction of arrival (DOA) estimation. The source number estimation is a very important part of DOA estimation, otherwise the incorrect estimation will lead to the deviation of the result[1][2].

At present, the common methods for source number estimation include Akaike information criterion (AIC) method based on information theory, minimum description length (MDL) method, method based on Gerschgorin disk theorem (GDE), and some methods based on clustering algorithm. Though the AIC and MDL methods are widely used[3], the estimation performance will be significantly worse with the decrease of signal-to-noise ratio (SNR), and the method is not suitable for the source number estimation of colored noise background[4][5]; GDE method estimates the source number by unitary transformation of the covariance matrix of the received data and by judging whether the radius of disk corresponding to the noise tends to zero[6][7]. This method is very dependent on the threshold, and is prone to estimation errors with low SNR.

Clustering analysis is a kind of unsupervised machine learning algorithm, which clusters the samples by similarity or close relationship among the data samples, and is not sensitive to colored noise. K-means algorithm is a commonly-used clustering algorithm[8] to estimate the number of signal sources at present. In addition, the algorithm based on fuzzy-clustering also can get high estimation accuracy, but it uses multiple features and extends the operation to multi-dimensional space. When the number of feature vectors increases, the amount of calculation will be very large[9][10]. After analyzing the defects of classical K-means clustering algorithm, this paper proposes a density-based clustering algorithm[11] to estimate source number. Then the detailed steps of ordering points to identify the clustering structure (OPTICS) algorithm[12] to estimate source number are also proposed. The experimental results show that the method based on OPTICS algorithm has better performance on source number estimation over others under the same conditions.
2. Array signal processing model

Taking the uniform linear array as an example, it is assumed that the elements are isotropic, the number of elements is \( N \), and the spacing is \( d \). There are \( D \) narrow-band signal sources incident on the antenna array in the form of plane waves in the far field, and the incident angle of the source is \( \theta_i \). The received vector by the antenna is

\[
X(t) = AS(t) + N(t)
\]

Where \( X(t) \) is the received vector, \( A = [a(\theta_1), a(\theta_2), ..., a(\theta_D)] \) is the array manifold matrix, \( S(t) \) is the signal vector and \( N(t) \) is the noise vector. The steering vector \( a(\theta_i) \) corresponding to the source \( i \) is the Vandermonde vector:

\[
a(\theta_i) = [1, e^{-j\frac{2\pi}{\lambda_0}d\sin(\theta_i)}, ..., e^{-j(N-1)\frac{2\pi}{\lambda_0}d\sin(\theta_i)}]^T
\]

The covariance matrix \( R \) signal received by the array is

\[
R = ARR_s^sA^H + R_N
\]

Where \( R_s \) is the source covariance matrix and \( R_N \) is the noise covariance matrix. When the signal is incoherent, the covariance matrix \( R \) is full rank and it has \( N \) eigenvalues \( \lambda_1, \lambda_2, \lambda_3, ..., \lambda_N \), corresponding to the eigenvector \( v_1, v_2, v_3, ..., v_N \). When the noise is Gaussian white noise, the noise variance is \( \sigma^2 \), and the distribution of eigenvalues is:

\[
\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_D \geq \lambda_{D+1} = \cdots = \lambda_N = \sigma^2
\]

In practice, the covariance matrix \( R \) is the autocorrelation matrix of the received sampling signal, i.e.

\[
R_{xx} = \frac{1}{n-1}\sum_{t=1}^{n} X(t)X^T(t)
\]

Due to the influence of noise and limited data length, the distribution of these eigenvalues is as follows:

\[
\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \hat{\lambda}_3 \geq \cdots \geq \hat{\lambda}_D \geq \hat{\lambda}_{D+1} \geq \cdots \geq \hat{\lambda}_N > \sigma^2
\]

Based on the above model, various methods can be used to estimate source number according to the features of eigenvectors and eigenvalues.

3. Defect analysis of K-means algorithm

Because K-means algorithm has better estimation performance with low SNR compared with MDL, GDE, and also prefers to AIC with high SNR, and it is not sensitive to color noise, it has been widely used. For the source number estimation, K-means algorithm divides the eigenvalues into noise cluster and signal cluster (\( k = 2 \)). The initial clustering centers are usually the minimum and maximum eigenvalues. The steps of K-means algorithm for source number estimation are as follows:

Step 1, select \( N_x \) snapshots of received signal as samples \( X(t) \) and calculate the autocorrelation matrix \( R_{xx} \);

Step 2, decompose \( R_{xx} \) to get \( N \) eigenvalues;

Step 3, select \( \lambda_{\text{min}} \) and \( \lambda_{\text{max}} \) as the clustering centers of the initial noise cluster and signal cluster;

Step 4, calculate the distance between eigenvalue \( \lambda_i \) and two clustering centers \( d_1, d_2 \). If \( d_1 \) is less than \( d_2 \), assign \( \lambda_i \) to the noise cluster, then update the mean value of the elements in the latest cluster to the noise cluster center. Otherwise, assign \( \lambda_i \) to the signal cluster and update the signal cluster center in the same way;

Step 5, review other eigenvalues until all eigenvalues no longer update their clusters. The number of elements in the signal cluster is the source number.

Although K-means algorithm has good estimation accuracy, the performance will be significantly worse when the SNR is lower than a certain threshold. Because of the ambiguity of signal and noise in low SNR and the algorithm relying on the initial clustering center, the method above easily causes the estimation errors under the low SNR condition. The following table shows a set of eigenvalues obtained by array with 20 elements under SNR=-7dB, while the source number is 2 (in order to facilitate the analysis, they have been arranged in the order of small to large), and the data is plotted in Figure 1. Obviously, No. 18 eigenvalue is more suitable to be clustered into noise cluster. However, K-means algorithm finally updates the clustering centers of the two clusters to 9.674 and 29.807, and the No. 18
eigenvalue is closer to the signal cluster center, so it is clustered into the signal cluster, resulting in the wrong source number estimation as 3.

Table 1. A set of eigenvalues of $R_{xx}$ at SNR=-7dB.

| No. | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|     | 3.504088 | 4.369283 | 5.244569 | 5.749 | 6.35647 | 6.935689 | 7.972339 | 8.194335 | 9.216038 | 9.78556 |
| No. | 11  | 12  | 13  | 14  | 15  | 16  | 17  | 18  | 19  | 20  |
|     | 10.31164 | 11.26302 | 12.00643 | 14.31024 | 14.84154 | 16.39874 | 18.0121 | 20.41598 | 32.44913 | 36.55609 |

Figure 1. Plotting the eigenvalues in table 1.

4. OPTICS algorithm for source number estimation

Based on the observation of eigenvalues above, it is more suitable to cluster with the “density” feature.

OPTICS algorithm is an improvement on density-based spatial clustering of applications with noise (DBSCAN) algorithm. The basic concept of DBSCAN algorithm is: taking each data as the centre, drawing a circle with radius $\varepsilon$ and giving the threshold value MinPts, the centre of the circle is a core point when the number of points in the circle is greater than or equal to MinPts, otherwise it is a low-density point. Two core points are connected when they are belonged to the same circle. The low-density points in the circle of the connected core points are called boundary points. The connected core points and their boundary points are connected to form a cluster. The algorithm is as follows:

When DBSCAN algorithm is used to estimate source number, the distance between eigenvalue $\hat{\lambda}_i$ and other eigenvalues is calculated, and appropriate $\varepsilon$ and MinPts are selected to obtain all the core points and their boundary points. The core points with a distance less than or equal to $\varepsilon$ and their boundary points are clustered into one cluster. The source number is $N$ minus the number of elements of cluster which contains $\hat{\lambda}_i$. For example, taking MinPts=2 and $\varepsilon=5$, the No.18 eigenvalue in chapter 3 will be clustered into the noise cluster. Obviously, DBSCAN algorithm is more suitable for source number estimation than K-means. However, the defect of DBSCAN algorithm is also obvious: the key parameters $\varepsilon$ and MinPts need to be set manually, which is not practical for blind signal estimation.

Based on the DBSCAN algorithm, the core distance and reachable distance are introduced in OPTICS algorithm. The core distance refers to the minimum radius that makes $x(x \in X)$ the core point for given $\varepsilon$ and MinPts. The expression is as follows:

$$CD(x) = \begin{cases} \text{undefined}, & |N_e(x)| < \text{MinPts} \\ d(x, N_e^{\text{MinPts}}(x)), & |N_e(x)| \geq \text{MinPts} \end{cases}$$

Where $N_e^i(x)$ is the data adjacent to the $i$th of $x$ in set $N_e(x)$. The reachable distance of $y$ about $x (x, y \in X)$ refers to:

$$RD(x) = \begin{cases} \text{undefined}, & |N_e(x)| < \text{MinPts} \\ \text{max} \{cd(x), d(x, y)\}, & |N_e(x)| \geq \text{MinPts} \end{cases}$$

Where $d(x, y)$ is the Euclidean distance between $x$ and $y$.

The goal of the OPTICS algorithm is to output an ordered sequence of objects. Each object in the sequence has two attributes: core distance and reachable distance. The reachable graph can be drawn by...
the reachable distance attribute of the object, and the number of elements of each cluster can be obtained intuitively. The steps of OPTICS algorithm for source number estimation are as follows:

Step 1, select \( N_x \) snapshots of received signal as samples \( \mathbf{X}(t) \) and calculate the autocorrelation matrix \( \mathbf{R}_{\mathbf{X}^*} \);

Step 2, decompose \( \mathbf{R}_{\mathbf{X}^*} \) to get \( N \) eigenvalues, and normalize them;

Step 3, calculate core distances of all eigenvalues with \( \varepsilon \) and \( \text{MinPts} \);

Step 4, select one of the eigenvalues as the inspection point, calculate its Euclidean distance \( d \) from other eigenvalues, compare \( d \) with its core distance, and replace the larger one as the reachable distance;

Step 5, take the closest object to the current inspection point as the new inspection point, repeat step 4, and finally get the inspection points arranged in order and their reachable distances;

Step 6, take the ordered number of \( k \) in \( \max [RD(k+1) - RD(k)] \), and the source number is estimated to be \( N-k \).

The data in Table 1 are clustered by using OPTICS algorithm, and the \( \varepsilon \) is taken as infinity, the \( \text{MinPts} \) is taken as 2, 3 or 4, to obtain the reachable graph by taking the no.7 eigenvalue as the initial inspection point (Figure 3). For the binary clustering problem, the number of sources can be estimated to be 2 correctly by getting the steepest value jump in the reachable graph. At the same time, it can be seen that the algorithm is not sensitive to \( \varepsilon \) and not strict to \( \text{MinPts} \) value.

![Figure 3. Reachable graph of eigenvalues in Table 1 obtained by OPTICS algorithm.](image)

5. Experiments and performance analysis

5.1. The effect of SNR on the estimation performance of different algorithms

In the experiment, two far-field narrow-band signals received by the antenna array come from the directions of 5° and 60°. The number of snapshots is 100, and the SNR increases from -15dB to 15dB at a step of 1dB. Through 200 Monte-Carlo experiments at each SNR, the curve of accuracy of source number estimation with SNR is shown in Figure 4. Under the same conditions, the source number estimation based on OPTICS still has more than 85% accuracy when the SNR is -9dB, and about 20% accuracy at a lower SNR, which is obviously better than the accuracy of K-means, MDL and GDE algorithm, and also better than AIC algorithm when the SNR is high.

5.2. The effect of snapshots on the estimation performance of different algorithms

In the experiment, the antenna array also receives two far-field narrow-band signals from the directions of 5° and 60°. The SNR is 10dB, and the number of snapshots increases from 10 to 200 at a step of 10. 200 Monte-Carlo experiments are conducted under each number of snapshots. The curve of accuracy of source number estimation with snapshots is shown in Figure 5. Under the same conditions, when the number of snapshots is greater than 20, the estimation accuracy of OPTICS and MDL is close to 100%, while K-means and GDE need more than 50 snapshots to achieve the same estimation effect. Because AIC is not an unbiased estimation algorithm, there is still a large error probability when the number of snapshots is large. The results show that the OPTICS algorithm has certain advantages in snapshots to achieve the same estimation accuracy.
6. Conclusion

Applied with features of covariance matrix eigenvalues of received data, a method based on OPTICS algorithm to estimate source number is proposed in this paper. The experimental results show that using OPTICS algorithm can overcome the inherent defect of K-means algorithm to estimate source number, and also has better performance over other algorithms under the same conditions. The estimation accuracy of OPTICS algorithm is improved in low SNR compared with others, but how to get higher accuracy in the case of lower SNR using fewer snapshots needs further study.

References

[1] Kumarean R O and Tufts D W 1983 Estimating the angle of arrival of multiple plane waves *IEEE Trans.on Aerospace Electron Systems* 19(1) 134-39.
[2] Schemidt R O 1986 Multiple emitter location and signal parameter estimation *IEEE Trans.on Antennas and Propagation* 34(3) 276-80.
[3] Wax M and Kailath T 1985 Detection of signals by information theoretic criteria *IEEE Trans.on Acoust, Speech, and Signal Processing* 33(2) 387-92.
[4] Fishler E and Messer H 2000 On the use of order statistics for improved detection of signals by the MDL criterion *IEEE Trans.on Signal Processing* 48(8) 2242-47.
[5] Zhang Q T and Wang K M 1993 Information theoretic criteria for the determination of the number of signals in spatially correlated noise *IEEE Trans.on Signal Processing* 41(4) 1652-63.
[6] Wu H T, Yang J F and Chen F K 1995 Source number estimation using transformed Gerschgorin Radii *IEEE Trans.on Signal Processing* 43(6) 1325-33.
[7] Caspary O, Nus P and Cecchin T 1998 The source number estimation based on Gerschgorin radii *IEEE Trans.on Acoust, Speech, and Signal Processing* 4 1993-96.
[8] Hartigan J A and Wong M A 1979 A K-means clustering algorithm *J Roy Stat Soc* 28(1) 100-08.
[9] Bao Zhiqiang, Han Bing and Wu Shunjun 2006 A new source number detection algorithm based on fuzzy clustering *J Electr Inf Technol* 28 1761-65.
[10] Hu Jun 2008 Research, simulation and application of signal number estimator based on fuzzy clustering *Journal of system simulation* 20(5) 1133-38.
[11] Ester M, Kriegel H P, Sander J and Xu Xiaowei 1996 A density-based algorithm for discovering clusters in large spatial databases with noise *KDD* 96 226-31.
[12] Ankerst M, Breunig M M, Kriegel H P and Sander J 1999 OPTICS:ordering points to identify the clustering structure *ACM SIGMOD Record* 28(2) 49-60.