Active source selection using Gap statistics for underdetermined blind source separation

ACTIVE SOURCE SELECTION USING GAP STATISTICS FOR UNDERDETERMINED BLIND SOURCE SEPARATION

Yuhui Luo and Jonathon Chambers

Centre for Digital Signal Processing Research
School of Physical Science and Engineering, King's College London
Strand, London WC2R 2LS, United Kingdom, Email: Yuhui.Luo@kcl.ac.uk

ABSTRACT

We address the problem of automatically determining the number of active sources in underdetermined blind source separation (BSS). A time-frequency approach to underdetermined BSS is exploited to discriminate the time-frequency structure of the measured mixtures. To determine the number of active sources over an observation interval, an advanced clustering technique based on Gap statistics is proposed. Simulation studies are presented to support the proposed approach.

1. INTRODUCTION

The aim of blind source separation (BSS) is to extract the underlying sources from a set of linear mixtures, which is typically obtained by a sensor array, without explicit information describing the sources and the system channels. Two assumptions are conventionally required in BSS. One is the mutual independent assumption of the input sources, which enables the approach of independent component analysis. For the second assumption, it requires that the number of measurement sensors to be at least equal to the number of sources. For, otherwise, the problem will be ill-conditioned.

If, however, the system has more sources than sensors, the problem is termed underdetermined BSS. So far, most underdetermined BSS methods resort to various sparsity within the sources. For example, in [7], an algorithm is proposed based upon the idea of representing the observation data with an overcomplete dictionary. However, this sparsity assumption for the sources may not hold in many applications and performance of the algorithm thereby degrades. Assuming the mixing matrix is known, several methods are suggested in [6] to estimate the input sources under different probabilistic source sparsity models. Another route for underdetermined BSS is to exploit the information embedded in the time-frequency (t-f) domain, which is particularly useful for non-stationary sources. By exploiting the t-f regions at which only one source is present, the algorithm in [1] is able to cancel the contribution from one source. This result however is far from the goal of complete source separation. Assuming a certain structure of the mixing matrix and that the t-f representations of input sources do not overlap, an offline method for separating an arbitrary number of sources from two mixtures is proposed in [3]. Also under the assumption of t-f orthogonality, by classifying the principle eigen vectors of the t-f matrices, an alternative algorithm is proposed, [4]. However, this approach suffers from the identification of phantom sources because of the rough classification employed. Thresholding in the later stage therefore becomes crucial and its operation requires the number of active sources be known a priori. In this paper, by applying an advanced clustering technique that is based upon Gap statistics [5], the drawback of phantom sources is overcome. More importantly, the newly proposed algorithm is able to separate different sources even when the number of active sources is not known, which is very likely to be the scenario in practice.

This research was funded by QinetiQ under the United Kingdom Ministry of Defence Corporate Research Programme CISP

2. DATA MODEL AND THE SIGNAL T-F QUADRATIC REPRESENTATION

A multi-input and multi-output system with k input sources and m measurement sensors is assumed. The observation signals are modelled as the instantaneous mixtures of the input sources. The notations $(\cdot)^T$, $(\cdot)^H$ and $(\cdot)^*$ denote respectively the operations of transpose, conjugate transpose and complex conjugate. At discrete time instant t, the measurement signal $\mathbf{x}(t) = [x_1(t)\ x_2(t)\ \dots\ x_m(t)]^T$ is written as

$$\mathbf{x}(t) = A\mathbf{s}(t) + \mathbf{n}(t) \tag{1}$$

where $\mathbf{s}(t) = [s_1(t)\ s_2(t)\ \dots\ s_k(t)]^T$ is the source vector, A is an $m \times k$ mixing matrix and $\mathbf{n}(t)$ is the zero mean additive noise. An unmixing matrix B is introduced to attempt to restore the mutual independence property of the sources, which is lost after linear mixing. The unmixing output $\mathbf{y}(t)$

0-7803-7946-2/03/$17.00 ©2003 IEEE

is the estimator of the input sources subject to a scaling and permutational ambiguity.

$$\mathbf{y}(t) = B\mathbf{x}(t) \tag{2}$$

For the measurement signal from sensor-i, $x_i(t)$, the discrete-time form Cohen's class t-f representation is written as

$$D_{x_i x_i}(t,\ f) = \sum_{l=-\infty}^{\infty}\sum_{v=-\infty}^{\infty} \phi_{ii}(v,l) \times x_i(t+v-l)\, x_i^*(t+v+l)\, e^{-j4\pi fl} \tag{3}$$

where $\phi_{ii}(v,l)$ is the signal-independent kernel function. The t-f representation of the correlation between two sensor signals $x_i(t)$ and $x_j(t)$ is written as $D_{x_i x_j}(t,\ f) = \sum_l \sum_v \phi_{ij}(v,l)\, x_i(t+v-l)\, x_j^*(t+v+l)\, e^{-j4\pi fl}$. In the application of BSS, the quadratic t-f representation can be extended to accommodate vector signals, [2] and [4],

$$D_{\mathbf{xx}}(t,\ f) = \sum_{l=-\infty}^{\infty}\sum_{v=-\infty}^{\infty} \boldsymbol{\phi}(v,l) \times \mathbf{x}(t+v-l)\, \mathbf{x}^H(t+v+l)\, e^{-j4\pi fl} \tag{4}$$

where $\boldsymbol{\phi}(v,l)$ is a matrix whose $(i,\ j)^{th}$ entry $\phi_{ij}(v,l)$ is the kernel associated with the i^{th} and the j^{th} measurement sensor outputs. Similarly, denote $D_{\mathbf{ss}}(t,\ f)$ as the t-f representation matrix of the sources, i.e., $D_{\mathbf{ss}}(t,\ f) = \sum_l \sum_v \boldsymbol{\phi}(v,l)\, \mathbf{s}(t+v-l)\, \mathbf{s}^H(t+v+l)\, e^{-j4\pi fl}$. It is shown in [2] that the t-f representation of the received signal is related to that of the sources by

$$D_{\mathbf{xx}}(t,\ f) = A D_{\mathbf{ss}}(t,\ f)\, A^H \tag{5}$$

This relationship will be further exploited in underdetermined BSS.

3. A CURRENT APPROACH AND ITS LIMITATION

In [4], the t-f supports of different sources are assumed not to overlap. Under this t-f orthogonality assumption, only one diagonal entry of $D_{\mathbf{xx}}(t,\ f)$ is different from zero. Specifically, if at $(t_i,\ f_i)$ only source-j is presented in the system, the matrix $D_{\mathbf{xx}}(t,\ f)$ in eqn. (5) is written as

$$D_{\mathbf{xx}}(t_i,\ f_i) = D_{s_j s_j}(t_i,\ f_i)\, \mathbf{a}_j \mathbf{a}_j^H \tag{6}$$

where $D_{s_j s_j}(t,\ f)$ is the t-f distribution of source-j at $(t_i,\ f_i)$ and $\mathbf{a}_j$ is the j^{th} column of the mixing matrix A, i.e., steering vector of source-j. As $D_{s_j s_j}(t_i,\ f_i)$ and $\mathbf{a}_j$ are effectively the principal eigen value and eigen vector of $D_{\mathbf{xx}}(t_i,\ f_i)$, eigen decomposition of the matrix $D_{\mathbf{xx}}(t,\ f)$ provides an estimator of the t-f representation of the input sources. Meanwhile, notice that for t-f points that are associated with the presence of the same source, in principle, the eigenvector of $D_{\mathbf{xx}}(t,\ f)$ should remain the same. This finding suggests that the signal spectra can be separated by clustering the t-f points that are associated to the same principal eigen vector. The source signals can then be synthesized with the estimator of $D_{s_j s_j}(t, f)$.

A major problem of the above approach is in the classification of the eigenvectors. In [4], two normalized vectors $\widetilde{\mathbf{a}}_j$ and $\widetilde{\mathbf{a}}_{j'}$ are allocated into two different classes if the angle between them is larger than a certain threshold, i.e.,

$$\arccos\left(\widetilde{\mathbf{a}}_j^T \widetilde{\mathbf{a}}_{j'}\right) > \varepsilon \tag{7}$$

where $\widetilde{\mathbf{a}}_j = \left[\mathrm{Re}(\mathbf{a}_j)^T\ \mathrm{Im}(\mathbf{a}_j)^T\right]^T$ and $\|\widetilde{\mathbf{a}}_j\|_2^2 = 1$. However, this method often gives more classes than the number of sources, as indicated by the authors therein and demonstrated by the simulations in later section. To remove artificial sources, further thresholding is required and in its operation the number of active sources, k, is assumed to be known a priori. It should be addressed that the assumption of priori knowledge of number of sources may not be satisfied in many applications and therefore a more advanced classification method is needed.

4. THE INTRODUCTION OF GAP STATISTICS IN UNDERDETERMINED BSS

The problem of classification refers to the issue of detecting the hidden structure in some data set Z. In the application of underdetermined BSS, the data set Z is the collection of eigenvectors $\widetilde{\mathbf{a}}_j$ and we expect it to be partitioned into k clusters, as there are k active sources in the system. Various clustering methods have been proposed, most of which assume that the number of clusters is known a priori. Let the well-known k-means algorithm be an example. In the k-means algorithm, a number k of centroids are initialized as vectors far apart. Then each member in Z is examined and assigned to one of the clusters depending on the minimum Euclidean distance. The positions of centroids are updated everytime a member is added to the cluster. Such a procedure continues until all the members are partitioned into k clusters. Although the k-means algorithm assumes the a priori knowledge of the number of clusters in the data set, which contradicts our ultimate goal of automatic determining the number of sources, various simulations conducted show that the k-means algorithm still outperforms the method in [4], i.e., eqn. (7). Source separation is successfully achieved without the generation of phantom sources.

When the value of k is not known, however, only a few methods are available in the open literature. The algorithm based on Gap statistics is one of the effective approaches. The identification of value k is accomplished by

noticing that the error measure decreases monotonically as the number of clusters increases. Specifically, the closeness between two vectors $\widetilde{\mathbf{a}}_j$ and $\widetilde{\mathbf{a}}_{j'}$ can be measured by squared Euclidean distance $\|\widetilde{\mathbf{a}}_j - \widetilde{\mathbf{a}}_{j'}\|_2^2$. Assume the data set is partitioned into k clusters $C_1, C_2, \ldots C_k$, where C_i is the set containing the member indices of the i^{th} cluster. The pairwise distances within the cluster C_i is given by $\sum_{j,j' \in C_i} \|\widetilde{\mathbf{a}}_j - \widetilde{\mathbf{a}}_{j'}\|_2^2$. Taking all the k clusters into account, the error measure in Gap statistics is defined as

$$W_k = \sum_{i=1}^{k} \frac{1}{2n_i} \sum_{j,j' \in C_i} \|\widetilde{\mathbf{a}}_j - \widetilde{\mathbf{a}}_{j'}\|_2^2 \qquad (8)$$

where n_i is the number of vectors within the i^{th} cluster. The logarithm of the above error measure, $log\,(W_k)$, is compared with that computed from a reference data set Z' which is drawn from an appropriate distribution. The natural number of clusters is then estimated as the value at which $log\,(W_k)$ falls the farthest below the reference curve, as the word 'Gap' in the name suggests.

It is clear that the generation of the reference distribution is crucial in the Gap test. If the dimension of $\widetilde{\mathbf{a}}_j$ is 1, the reference distribution can be chosen as the uniform distribution, [5], since such a distribution is most likely to produce fake clusters. Mathematically, the following equality exists for $k \geq 1$

$$\inf\left\{\frac{MSE_Z(k)}{MSE_Z(1)}\right\} = \frac{MSE_U(k)}{MSE_U(1)} \qquad (9)$$

where inf $\{\cdot\}$ denotes the infimum operator, i.e., the maximum lower bound.

In a more general case, the dimension of the data in Z is larger than unity. In other words, the data in Z should be characterized by more than one feature. No distribution can satisfy eqn (9) unless its support is degenerate to a subset of a line. Two ways are proposed for the generation of a reference distribution. The first one is straight forward. Each reference feature is generated uniformly over the range of the observed values corresponding to that feature. In comparison, the reference data Z' in the second method is drawn from a uniform distribution according to the direction of the principal component of the data Z, [5]. Specifically, the data Z are transformed by $N = ZV$, where V is obtained from the singular value decomposition of Z, i.e., $Z = UDV^T$. Then uniform distributed data N' is generated over the range of the columns of N. Finally, the reference data Z' is given by the reverse-transform $Z' = N'V^T$. Simulation studies suggest that the second method is more robust than the first one in various scenarios. Employing the k-means algorithm, the method of Gap statistics can be summarized into three steps, [5].

(1) The observed data Z are clustered by varying the total number of clusters from $k = 1, 2 \ldots K$. Compute the error measure $W_k, k = 1, 2, \ldots K$.

(2) Generate a number, L, of reference data sets using one of the methods mentioned above and cluster each one with the k-means algorithm. Compute the dispersion measure, W'_{kl}, $l = 1, 2, \ldots L$ and $k = 1, 2, \ldots K$. The Gap statistics are then given by

$$Gap(k) = \frac{1}{L}\sum_{l} \log(W'_{kl}) - \log(W_k) \qquad (10)$$

(3) To account for the sample error in approximating an ensemble average with L reference distributions, the standard deviation is computed as $sd_k = \left(\frac{1}{L}\sum_l \left(\log(W'_{kl}) - \bar{l}\right)^2\right)^{1/2}$, where $\bar{l} = \frac{1}{L}\sum_l \log(W'_{kl})$. Its influence on the Gap test is given by $sd'_k = sd_k \cdot \left(1 + \frac{1}{L}\right)^{1/2}$. The number of clusters is estimated as the smallest k such that

$$Gap(k) \geq Gap(k+1) - sd'_{k+1} \qquad (11)$$

Upon the introduction of Gap statistics in underdetermined BSS, not only the drawback of phantom sources in [4] is overcome but also the t-f separation algorithm is now operating in a completely blind fashion. The requirement of knowing the number of sources is eliminated. The disadvantage of the proposed approach is in the additional computational complexity needed.

5. SIMULATION

We assume a three sources and two sensors system. The test signals are assumed to be the linear chirp signals, which are orthogonal in the t-f plane. The kernel function is selected as the Choi-Williams distribution. Assume the signal to noise ratio is 20dB and the mixing matrix is $A = \begin{bmatrix} 0.4+0.7i & 0.9+0.2i & 0.6+0.5i \\ 0.6+0.6i & 0.7+0.2i & 0.4+0.3i \end{bmatrix}$. The source signals and observation mixtures are shown in Fig 1 (a)-(e). With the classification method in [4], the number of clusters found is summarized in Table 1. When the threshold $\varepsilon = 0.4$, three clusters have been found. But the sources are still mixed in the t-f domain. If ε is slightly reduced to 0.3, even though one phantom source has already been generated, source separation still fails, as shown in Fig 1 (f)-(i). Only when the threshold value goes down to 0.1, can successful source separation be achieved. However, 12 in number phantom sources will be generated. In the experiment for Gap statistics, 15 ($L = 15$) copies of reference data are generated. The logarithm of the error measure obtained from the actual data set and the reference data set is described in Fig 2(a) and the result of Gap test versus different numbers of clusters is shown in Fig 2 (b). The peak at the value of three in Fig 2 (b) indicates three natural clusters in the data set, which is a desirable result. The t-f distribution of the three sources are successfully separated, as confirmed in Fig 3.

Threshold ε	0.4	0.3	0.2	0.1
No. of cluster found	3	4	7	15
Successful Separation (Y/N)	N	N	N	Y

Table 1. Number of clusters found with the rough classification method

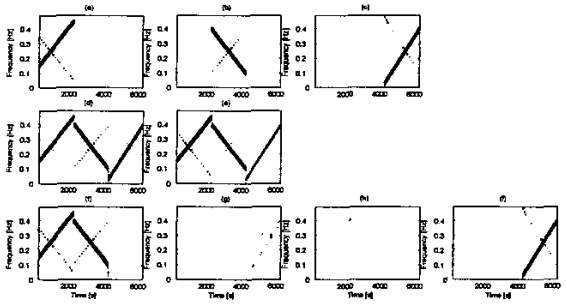

Fig. 1. A three sources and two sensors system. (a) source-1 (b) source-2 (c) source-3 (d) sensor-1 (e) sensor-2 (f) unmixing output-1 (fails) (g) unmixing output-2 (fails) (h) unmixing output-3 (fails) (g) unmixing output-4 (successful)

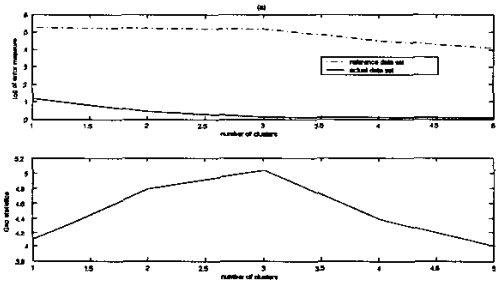

Fig. 2. Gap statistics in underdetermined BSS (a) Logarithm of the above error measure obtain from the actual and reference data set (b) Gap statistics

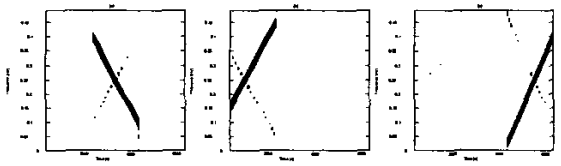

Fig. 3. Successful underdetermined BSS with the Gap statistics (a) unmixing output-1 (b) unmixing output-2 (c) unmixing output-3

6. CONCLUSION

When separating more sources than observed mixtures on the basis of temporal sparsity of the sources, the time-frequency BSS method in [4] suffers from the problem of phantom sources and requires knowledge of the number of active sources. By employing the method of Gap statistics, a solution has been proposed in this paper to automate the selection of the number of active sources. Simulations show that the proposed solution tackles the problem of underdetermined BSS when the number of active sources is unknown.

7. REFERENCES

[1] F. Abranrd, Y. Deville, and P. White. From blind source separation to blind source cancellation in the underdetermined case: A new approach based on time-frequency analysis. *Third Inter. Conf. on Independent Component Analysis and Blind Signal Separation*, 2001.

[2] A. Belouchrani and M. G. Amin. Blind source separation based on time-frequency signal representation. *IEEE Trans. on Signal Processing*, 46(11):2888–2897, 1998.

[3] A. Jourjine, S. Rickard, and O. Yilmaz. Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures. *Proc of IEEE Int. Conf. Acoust. Speech Signal Processing*, 5:2985–2988, 2000.

[4] L.-T. Nguyen, A. Belouchrain, K. Abed-Meraim, and B. Boashash. Separating more sources than sensors using time-frequency distribution. *Sixth International Symposium on Signal Processing and its Applications*, 2:583–586, 2001.

[5] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a dataset via the gap statistic. *Tech. report, Stanford University, Published in JRSSB*, 2000.

[6] L. Vielva, D. Erdogmus, and J. C. Principe. Underdetermined blind source separation using a probabilistic source sparsity model. *Third Inter. Conf. on Independent Component Analysis and Blind Signal Separation*, 2001.

[7] M. Zibulevsky and B. A. Pearlmutter. Blind source separation by sparse decomposition in signal dictionary. *Neural Computation*, 13(4):863–882, 2001.