MLSP 2005 Competition:
LARGE SCALE, ILL-CONDITIONED BLIND SOURCE SEPARATION PROBLEM with LIMITED NUMBER of SAMPLES

Andrzej CICHOCKI
Riken, Brain Science Institute,
Laboratory for Advanced Brain Signal Processing,
Wako-Shi, JAPAN

Deniz ERDOGMUS
Department of Computer Science and Engineering,
OGI School of Science and Engineering,
Oregon Health & Science University, USA

We assume a linear standard model described in matrix form as:

$$\mathbf{X} = \mathbf{A}\,\mathbf{S},$$

where **X** is an *n* by *T* matrix representing *n* observations for *T* consecutive time instants, **A** is an *n* by *n* nonsingular mixing matrix and **S** is an *n* by *T* matrix representing sources; *n* is the number of sources (equal to the number of sensors) and *T* is the number of available samples.

It is assumed that only the matrix **X** is available, while the matrices **A** and **S** are unknown and should be estimated. The objective of this problem is to investigate the effect of increasing dimensionality *n*, decreasing number of samples *T*, increasing ill-conditioning of the mixing matrix **A** and/or increasing the level of additive noise in the sensor level for performance and reliability of blind source separation algorithms.

The original source signals are to be generated in MATLAB as follows:

Set 1: random non-negative source signals:

S1=rand(n,T); (sub-Gaussian)

and

S2= -log(rand(n,T)).*max(0,sign(rand(n,T)-0.5)); (super-Gaussian)

It is assumed that the matrix of original sources **S** is build up in such a way that **t**hat approximately 50% of the sources are taken form the subset **S1** and 50% from the subset **S2**, the maximum number of available samples is *T*=5000, that is for the even number of sources they should be generated as follows:

S=[S1;S2];

Set 2: Source signals with temporal structure **SP** (natural 600 speech signals in Polish language sampled at 4 kHz) for *T*=20000 are available on the web pages (generated by Dr. Tomek Rutkowski):
http://www.bsp.brain.riken.jp/data/speechPOLISH_random_4kHz_200sources.mat

Only maximum *T*=5000 samples can be used using for any time sub-window.

The problem consists of four sub-problems.

1. Large scale problem

The dimension of the randomly generated mixing matrix **A**=rand(*n*) increases. The number of sources also correspondingly increases until limit is achieved with respect to reliable performance of the estimated mixing matrix. The sources have fixed number of samples (maximum *T*=5000). Acceptable performance is assumed that overall separation signal-to-interference ratio defined below is SIR > 15 dB.

The winner in this sub-category will be a person or a team who implements or develops any algorithm or set of algorithms (not necessarily a new algorithm) that separates the largest number of sources reliably. The overall SIR is calculated in MATLAB using:

```
G=W*A;
SIR=mean(10*log10(max(G.^2,[],2)./(sum(G.*G,2)-max(G.^2,[],2))));
```

and should be better (larger) than 15 dB. Here **W** is the separation matrix such that **Y=WX** are the independent sources.

Since the SIR is global performance index which does not take into account a quality of estimation of each individual source a cross-talking for each estimated source should be also below some threshold, specified here as less than 15 %.

2. Incomplete or reduced set data problem: X=AS

The number of sources is fixed, *n*=50, **A**=rand(50). However, the number of available samples is gradually reduced from *T*=5000.

The winner in this sub-category will be the person or team who implements or develops an algorithm which could estimate approximately (but reliably) the randomly generated matrices **A** for the lowest possible number of samples (neglected scaling and permutation of columns) with SIR >15dB.

3. Very ill-conditioned problem

Ill-conditioning and the dimension of the problem **X=AS** is gradually increased with the mixing matrix chosen as the Hilbert matrix **A**=**H**=hilb(*n*), where the dimension *n* is increased, starting from *n*=5.

The winner in this sub-category will be the person/team who proposes the algorithm (not necessary a new algorithm), which separates reliably the largest number of sources with the overall signal-to-interference ratio (SIR>15 dB). In other words, as stopping criterion is assumed that the SIR is not less than 15 dB.

Remark: It is assumed that matrices **H** and **S** are unknown. Moreover, the structure of the matrix **H** is also assumed to be unknown, i.e., the proposed algorithm should work for other very ill conditioned mixing matrices, although the challenge is to use the Hilbert matrix.

4. Noisy problem

Suppose that **X=AS+N,** where **N** is a n by T matrix representing additive Gaussian or uniform distributed noise with increasing noise level with respect to unit variance source signals, i.e., decreasing SNR (signal to noise ratio) from SNR= 20dB. The number of sources n=50 and the number of samples T=5000 are fixed. The mixing matrix is randomly generated as **A**=rand(n) with the condition number cond(**A**)<100 (so that the problem is not ill-conditioned).

The winner in this sub-category will be the person/team who estimates approximately the randomly generated matrices **A** for the lowest possible SNR for SIR > 15 dB. Notice that this measure does not consider the noise corruption levels at the separated outputs, rather it is only concerned with the performance of the (inverse)model estimation.

Remarks:

1. Sources **S1** and **S2** are non-negative so alternative methods to ICA such as NMF (Non-negative Matrix Factorization) or ICA with non-negativity constraints can be also implemented and tested.

2. The second set of provided speech sources **SP** have temporal structures, therefore second-order statistics (SOS) methods can be employed including also the SCA (sparse component analysis) in the time-frequency domain.

3. All sub-problems should be attempted using at least one subset of sources.

4. The proposed algorithms should solve any sub-problem in reasonable computation time, say, in less than one hour on a typical PC.

5. All experiments should be conducted in a Monte Carlo fashion with at least 100 independent runs. The performance bounds below (SIR>15dB) should be satisfied 90% of the time in these Monte Carlo simulations.

6. Report the results in a document (*.doc or *.pdf) where a brief description of the algorithm and appropriate references as well as experimental results demonstrating performance on the selected sub-problems are included. In your submission, please also include a self-contained Matlab script of your code to facilitate the replication of results if required.

Please note that the MLSP Committee has right to test the submitted MATLAB procedures and algorithms for their own generated data for the set **S1** or **S2** and the natural speech signals **SP**. *It is expected that participants investigate and submit results for at least two sub-problems.*

Regarding the performance index, the user can additionally use the recently provided MATLAB package BSS_EVAL BSS_EVAL is a MATLAB toolbox to compute reliably performance measures in (blind) source separation within an evaluation framework where the original sources are available as ground truth.
Download page:http://www.irisa.fr/metiss/bss_eval/