A NEW ALGORITHM IN BLIND SOURCE SEPARATION FOR HIGH DIMENSIONAL DATA SETS SUCH AS MEG DATA

[1] JALIL TAGHIA, [2] MOHAMMAD ALI DOOSTARI, [3] JALAL TAGHIA
[1,2] Dept. of Electrical Engineering, Shahed University, Tehran, IRAN
[3] Dept. of Electrical and Computer Engineering, Shahid Beheshti University, Tehran, IRAN

Abstract: - BSS is one of the well-known methods of signal processing. This method is based on recovering of original sources from observed mixtures without any further information about mixing system and original sources. In many applications, mixtures are combination of Non-Gaussian and Time-Correlated components. MCOMBI algorithm is known as a method for separation of these kinds of sources. The performance and accuracy of this algorithm is noticeable but the high computational cost is one of the most significant limitations of MCOMBI algorithm, especially for high dimensional datasets like high density Electroencephalographic (EEG) or Magnetoencephalographic (MEG) recordings. In this paper, we propose a new algorithm which uses combination of WASOBI and EFICA algorithms (like MCOMBI algorithm). In addition we use clustering method to decrease computational cost. In contrast with MCOMBI algorithm, the proposed algorithm decreases run time of separation and it has high accuracy close to MCOMBI algorithm. Thus, this algorithm is suitable for real high dimensional datasets. In this paper we use our algorithm for separation of artifacts in real MEG data.

Key-Words: - Statistical signal analysis, Biomedical signal processing, Blind source separation, Non-Gaussianity, Time-Correlation, MEG data, Independent component analysis

1 Introduction

Blind Source Separation (BSS) is one of the famous methods in the signal processing. This method estimates original sources only from the observed mixtures and separating operation performs without any prior information about the mixing system. In this paper we consider the most common BSS problem in which the sources are assumed to be independent and the mixing system is assumed to be linear and instantaneous. The BSS model can be expressed as:

$$x(t) = \sum_{j=1}^{d} a_j s_j(t) = As(t) \qquad (1)$$

where $A = [a_1, ..., a_d]$ is an unknown mixing matrix, $S(t) = [s_1(t), ..., s_d(t)]^T$ are the original unobserved sources and $X(t) = [x_1(t), ..., x_d(t)]^T$ are the observed linear and instantaneous mixtures. The BSS problem consists in estimating a separating matrix $\widehat{W} \approx A^{-1} = W$

practical limitation of all these combination approaches is that their computational cost is unaffordable for high-dimensional mixtures like the ones found in high-density electroencephalography (EEG) and magneto-encephalography (MEG). In this paper, we propose new algorithm that uses combination of WASOBI and EFICA algorithms (like MCOMBI algorithm). In addition we use clustering method to decrease computational cost. In contrast with MCOMBI algorithm, the proposed algorithm decreased run time of separation with high accuracy close to MCOMBI algorithm. We show that applying of this algorithm is suitable for real high dimensional datasets. In this paper we use this proposed algorithm for separation of artifacts in real MEG data.

2 Multidimensional Independent Components

Standard BSS assumes that the one-dimensional unknown sources in Eq.1 are mutually independent according to the independency contrast used. A generalization of this principle assumes that not all the d sources are mutually independent but they form M higher dimensional independent components [11], [12]. Let d_l denote the dimensionality of the l th multidimensional component that groups together the one-dimensional source signals with indexes $l_1,...,l_{d_l}$. Then, the l th multidimensional component is given by $S_l = [s_{l_1},...,s_{l_{d_l}}]^T$, where $l = 1,...,M$ and $d_1 + d_2 + ... + d_M = d$. Therefore, we can rewrite the sources data matrix S in Eq.1 as $S = [s_1,...,s_d]^T = Q[S_1,...,S_M]^T$ where Q is a permutation matrix. Using the above notation and dropping matrix Q under the permutation indeterminacy of ICA, we can reformulate Eq.1 as:

$$S = WX = [W_1X,...,W_MX]^T = [S_1,...,S_M]^T \quad (2)$$

The goal of multidimensional BSS is to estimate the sub-matrices $\{W_l\}_{l=1,...,M}$ each of which is of dimension $d_l \times d$. Since the sub-components of a multidimensional independent component are arbitrarily mixed, we can recover $\{W_l\}_{l=1,...,M}$ only up to an invertible matrix factor [12]. A multidimensional component according to certain independency contrast (e.g. Non-Gaussianity) might be separable into one-dimensional components using an alternative independency measure (e.g. cross-correlations). This suggests a procedure for combining complementary independency criteria [10]:

1. Try BSS using certain independency criterion.
2. Detect the presence of multidimensional components in the source signals estimated in step 1.
3. Try BSS using an alternative independency contrast in each multidimensional component found in step 2.

This is the basic idea underlying proposed algorithm which combines the complementary strengths of the Non-Gaussianity criterion of EFICA and the criterion based on cross-correlations of WASOBI.

3 Detection of Multidimensional Independent Components

A common way of evaluating the accuracy of the separation produced by any BSS algorithm is the matrix of Interference-to-Signal Ratios (ISR matrix). Element-wise, the ISR matrix is defined as $ISR_{kl} = G^2{}_{kl} / G^2{}_{kk}$ where $G = \hat{W}A$. $\hat{W}$ is the estimated separating matrix and A is the true mixing matrix. ISR_{kl} measures the level of residual interference between the k th and l th estimated components. The total ISR of the k th estimated source is defined as:

$$isr_k = \sum_{l=1, l\neq k}^{d} ISR_{kl}. \quad (3)$$

EFICA and WASOBI share the rare feature of allowing the estimation of the obtained ISR matrix through simple empirical estimate of $E[ISR]$ using the estimated sources $\hat{S}$. This means that EFICA and WASOBI permit us to estimate $\hat{ISR} \approx E[ISR]$. It has been shown that the estimations $\hat{ISR}$ obtained by WASOBI and EFICA are quite accurate even when the respective assumptions about the sources are only partially fulfilled [10]. The information provided by $\hat{ISR}$ is crucial for detecting the presence of multidimensional components within the estimated sources which is the reason for us to choose EFICA and WASOBI in our combined BSS method.

If the ISR matrix is known, or if it can be estimated, we can easily assess the presence of multidimensional independent components by grouping together components with high mutual interference. This is done by defining a symmetric distance measure between two estimated components as follows,

$$D(\hat{s}_k, \hat{s}_l) = D_{kl} = \frac{1}{ISR_{kl} + ISR_{lk}} \geq 0 \quad \forall \; l \neq k$$

$$D_{kk} = 0 \quad \forall \; k \quad (4)$$

Using the distance metric D, we cluster together the estimated components whose distance from each

other is small. For this task we use agglomerative hierarchical clustering [13] with single linkage. By single linkage we mean that the distance between clusters of components is defined as the distance between the closest pair of components. The output of this clustering algorithm is a set of $i = 1,...,d$ possible partition levels of the estimated sources. At each particular level the method joins together the two clusters from the previous level which are closest in distance. Therefore, in level $i = 1$ each source forms a cluster where as in level $i = d$ all the sources belong to the same cluster. For assessing the goodness-of-fit of the $i = 2,...,d-1$ partition levels, we propose using the validity index $I_i = D_i^{\text{int}era} / D_i^{\text{int}er}$ where $D_i^{\text{int}era}$ and $D_i^{\text{int}er}$ roughly measure, respectively, the average intra-cluster and inter-cluster distances. They are defined, for $1 < i < d$, as follows:

$$D_i^{\text{int}era} = \frac{\sum_{j=1,card(\Gamma_{i,j})>1}^{d-i+1} Card(\Gamma_{i,j})(Card(\Gamma_{i,j})-1)/2}{\sum_{j=1,card(\Gamma_{i,j})>1}^{d-i+1} \sum_{k\in\Gamma_{i,j},l\in\Gamma_{i,j}} ISR_{kl}}$$

$$D_i^{\text{int}er} = \frac{\sum_{j=1}^{d-i+1} Card(\Gamma_{i,j})(d - Card(\Gamma_{i,j}))}{\sum_{j=1}^{d-i+1} \sum_{k\in\Gamma_{i,j},l\notin\Gamma_{i,j}} ISR_{kl}} \quad (5)$$

where $\Gamma_{i,j}$ is the set of indexes of the sources belonging to the j th cluster at the i th partition level and $Card(\Gamma_{i,j})$ determines the number of elements in $\Gamma_{i,j}$. We also define $I_1 = 1/ISR_{\max}$ where $ISR_{\max}$ is the maximum entry in the ISR matrix. We set $I_d = 10$. Finally we choose the best cluster partition to be that one corresponding to the maximum of all local maxima of the cluster validity index I. By setting $I_d = 10$ we consider that the separation failed completely (there is just one d-dimensional cluster) if $D_i^{\text{int}er} < 10$, $D_i^{\text{int}era} \forall i = 2,...,d-1$. The definition $I_1 = 1/ISR_{\max}$ means that the estimated sources will be considered to be 1-dimensional (perfect separation) if $ISR_{\max} < \min_{i>2}(1/I_i)$. Therefore, since $I_d = 10$, we require the maximum ISR between two 1-dimensional components to be in any case below -10 dB. In order to ease the explanation of proposed algorithm in the next section we will use the following MATLAB notation to refer to the hierarchical clustering algorithm described in this section: $[i, I] = hclus(ISR)$ where the input parameter is the estimated ISR matrix, the first output parameter is the selected partition level and the second output parameter is a $1\times(d-i+1)$ cell array such that $I\{k\}$ is a vector containing the indexes of the sources belonging to the k th cluster.

4 Proposed Algorithm

The proposed algorithm is described using MATLAB notation as below:

```
function [B] = proposed-fun (X,ARorder)
[d, L] = size(X) ;
[B, ISRwa] = WASOBI (X,ARorder) ;
[iwa, Iwa] = hclus(ISRwa) ;
if iwa = = 1 , return; end
for i = 1: (d-iwa+1) ,
    if length (Iwa{i}) = = 1, continue; end
    index = Iwa{i}; di = length(index) ;
    [Bef, ISRef] = EFICA (B(index, :)*X) ;
    [ief, Ief] = hclus(ISRef) ;
    if (ief < di) || ...
      (min(sum(ISR(index,index),2)) > ...
       min (sum(ISRef,2))),
          B(index,:) = Bef*B(index,:) ;
    end
end
```

Proposed algorithm starts by applying WASOBI on the input data. The reason for using WASOBI first instead of EFICA is that the former is considerably faster than the latter for high dimensional mixtures, which is the target application of proposed algorithm. Subsequently, EFICA is applied on each multi-dimensional component of sources found in the output of WASOBI. Finally, we decide whether EFICA was able to improve the separation of the sources within the cluster or not. In our implementation of the algorithm we include a third step (not shown in the MATLAB code above) that consists on running WASOBI again on the cluster of unresolved components in the output of EFICA (if such a cluster exists). This last step is helpful only in the rare cases when, in the first run of WASOBI, we were not able to detect the correct clusters. If EFICA was able to separate some Non-Gaussian sources we expect the accuracy of WASOBI to improve by applying it only to the cluster of Gaussian components that was not correctly separated by EFICA. WASOBI requires the user to specify the order of the AR model that best fits the unobserved sources. However, the performance is not critically dependent on this parameter and it is enough to select an order high enough to model appropriately the source signals.

5 Separation via Proposed Algorithm

In this section we use proposed algorithm for separation of artifacts in MEG data. Moreover we compare speed and accuracy of proposed algorithm with MCOMBI algorithm.

5.1 Introducing Used MEG Data

Magnetoencephalography (MEG) is a noninvasive technique by which the activity or the cortical neurons can be measured with very good temporal resolution and moderate spatial resolution. When using a MEG record, as a research or clinical tool, the investigator may face a problem of extracting the essential features of the neuromagnetic signals in the presence of artifacts. The amplitude of the disturbances may be higher than that of the brain signals, and the artifacts may resemble pathological signals in shape.

In this paper we use proposed algorithm to separate brain activity from artifacts. The approach is based on the assumption that the brain activity and the artifacts (e.g. eye movements or blinks, or sensor malfunctions) are anatomically and physiologically separate processes and this separation is reflected in the statistical independence between the magnetic signals generated by those processes. The MEG signals were recorded in a magnetically shielded room with a 122-channel whole scalp Neuromag-122 neuromagnetometer. This device collects data at 61 locations over the scalp, using orthogonal double-loop pick-up coils that couple strongly to a local source just underneath. The test person was asked to blink and make horizontal saccades, in order to produce typical ocular (eye) artifacts. Moreover, to produce myographic (muscle) artifacts, the subject was asked to bite his teeth. Yet another artifact was created by placing a digital watch one meter away from the helmet into the shielded room. In Fig.1 we present a subset of 9 observed MEG signals from the frontal, temporal and occipital areas.

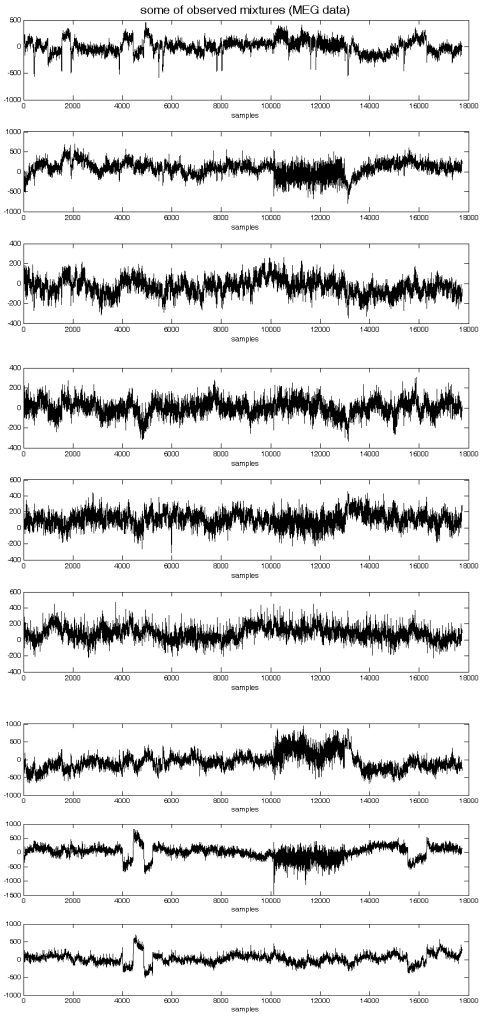

Fig.1: 9 observed signals from MEG data.

5.2 Experimental Results

We applied proposed algorithm on MEG data (introduced in last section) to separate artifacts from it. Fig.2 shows 8 estimated independent components (ICs) that found from the recorded data after applying proposed algorithm. The first two ICs (i.e. IC_1, IC_2) are clearly due to the muscular activity originated from the biting. IC_3 and IC_4 show the horizontal eye movements and the eye blinks, respectively. IC_5 represents of the cardiac artifact that is very clearly extracted. Sixth independent component (i.e. IC_6) is due to breathing. To find the remaining artifacts, the data were high-pass filtered, with cutoff frequency at 1Hz. Next, the independent component IC_7 was found. It shows clearly the artifact originated at the digital watch, near of the magnetometer. The last independent component IC_8 is related to a sensor presenting higher RMS (root mean squared) noise than the others. To evaluate the overall separation performance of proposed algorithm and MCOMBI, we used the average of the ISR obtained for the individual sources, i.e.:

$$ISR_{avg} = \frac{1}{d}\sum_{k=1}^{d} isr_k \qquad (6)$$

Where d indicates number of multidimensional components and isr_k can obtained from Eq.3. In Fig.3 we show the average Signal-to-Interference Ratio (SIR), obtained for different number of data samples of the sources, to compare accuracy of two algorithms. In this figure, dashed line (--) and solid line present SIR_{avg} which is computed from MCOMBI and proposed algorithm, respectively. According to Fig.3, the proposed algorithm is almost as accurate as MCOMBI algorithm.

On the other hand, the clustering method is used to decrease computational cost and running time of our

algorithm respect to MCOMBI algorithm. For comparing the speed of two algorithms, we have evaluated running time of them and presented results in Table 1. This can be understood from Table 1 that the computation time of applying our algorithm on MEG data is clearly smaller than the computation time of applying MCOMBI. The major advantage of our algorithm is the possibility of using it with very high dimensional datasets. It is noticeable that obtained running time is average of 10 times iteration of two algorithms individually and implementation of two algorithms is performed using MATLAB v7.0.4 software in computer with below system properties: Intel(R) Pentium(R) 4 CPU 1.70GHz / 512MB of RAM.

6 Conclusion

We proposed a new BSS algorithm that is combination of WASOBI and EFICA algorithms. This algorithm simultaneously separates Non-Gaussian and Time-Correlated sources. Moreover we used clustering method for decreasing computational cost and running time. We applied our algorithm and MCOMBI algorithm on the real MEG data with high dimensional data sets in order to separate artifacts from it. Furthermore, we demonstrated that proposed algorithm is almost as accurate as the MCOMBI algorithm. Moreover, because of using clustering method, our algorithm has lower computational cost respect to MCOMBI. Thus our algorithm is suitable for high dimensional data sets.

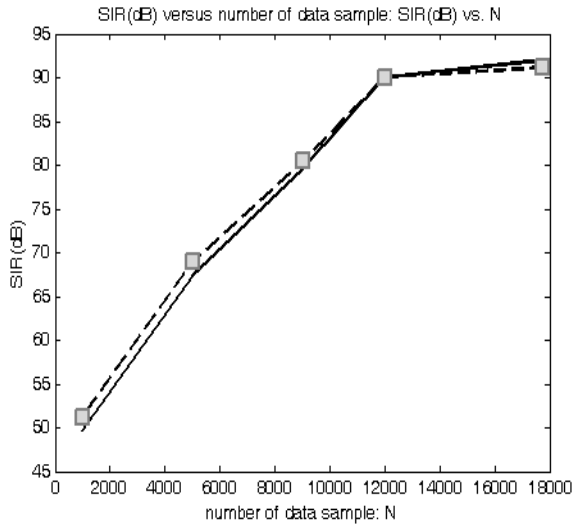

Fig.3: Average Signal-to-Interference Ratio (SIR_{avg}) obtained for different number of data samples of the sources. Dashed line (--) and solid line present SIR_{avg} that computed from MCOMBI and proposed algorithm, respectively.

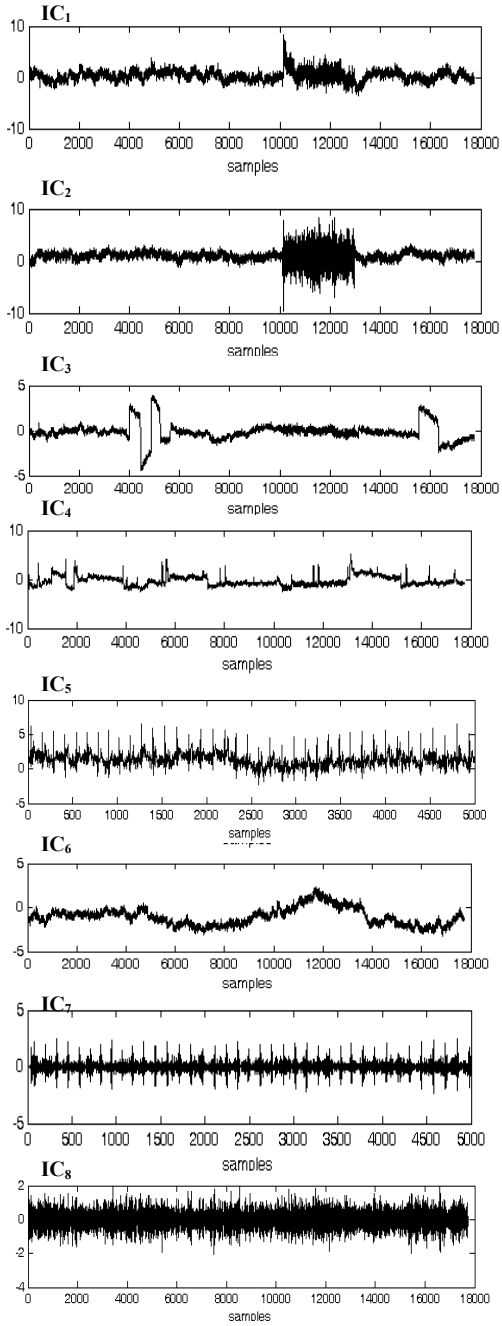

Fig.2: Independent components (Artifacts) estimated using proposed algorithm.

Table 1: Running time of MCOMBI and proposed algorithms.

Type of algorithm	Running time in seconds
MCOMBI	**330**
Proposed algorithm	**42**

Acknowledgements

We thank Doctor M. Kakooei and Pastor Hospital, for the MEG data used as example of a real biological environment.

References:

[1] A. Hyvärinen, Fast and robust fixed point algorithmms for Independent Component Analysis, *IEEE T. on Neural Networks*, vol.10, no. 3, 1999, pp. 626-634.

[2] J.F. Cardoso, High-order contrasts for Independent Component Analysis, *Neural Computation*, 1999, pp. 157-192.

[3] Z. Koldovský, P. Tichavský, and E. Oja, Efficient variant of algorithm Fastica for independent component analysis attaining the cramerrao lower bound, *IEEE T. Neural Networks*, vol.17, no.5, 2006, pp. 1265-1277.

[4] A. Belouchrani, K. Abed Meraim, J.F. Cardoso, and E. Moulines, A blind source separation technique based on second order statistics, *IEEE T. on Signal Processing*, vol.45, no.2, 1997, pp. 434-444.

[5] A. Ziehe and K.R. Müller, TDSEP-an efficient algorithm for blind separation using time structure, *In Proc. ICANN*, 1998, pp. 675-680.

[6] A. Yeredor, Blind separation of Gaussian sources via second-order statistics with asymptotically optimal weighting, *IEEE Signal Processing Letters*, vol.7, no.7, 2000, pp. 197-200.

[7] E. Doron and A. Yeredor, Asymptotically optimal blind separation of parametric Gaussian sources, *In Proc. ICA*, 2004, Granada, Spain.

[8] P. Tichavský, E. Doron, and A. Yeredor, A computationally affordable implementation of an asymptotically optimal BSS algorithm for AR sources, *In Proc. EUS-IPCO*, 2006, Florence, Italy.

[9] P. Tichavsk´y, Z. Koldovsk´y, E. Doron, A. Yeredor, and G. G´omez-Herrero, Blind signal separation by combining two ICA algorithms: HOS-based EFICA and time structure-based WASOBI, *In Proc. EUSIPCO*, 2006, Florence, Italy.

[10] P. Tichavsk´y, Z. Koldovsk´y, A. Yeredor, G. G´omez-Herrero, and E. Doron, A hybrid technique for blind separation of Non-Gaussian and time-correlated sources using a multi-component approach, *IEEE Trans. Neural Networks*, 2007.

[11] L. De Lathauwer, D. Callaerts, B. De Moor, and J. Vandewalle, Fetal electrocardiogram extraction by source subspace separation, *In Proc. IEEE Signal Processing workshop on higher-order statistics*, Girona, Spain, 1995, pp. 134-138.

[12] J.F. Cardoso, Multidimensional independent component analysis, *In Proc. ICASSP*, Seattle, WA, 1998.

[13] S. Winter, H. Sawada, S. Araki, and S. Makino, Hierarchical clustering applied to overcomplete BSS for convolutive mixtures, *Workshop on Statistical and Perceptual Audio Processing SAPA*, 2004, Korea.