# Blind Source Separation: New Tools for Extraction of Source Signals and Denoising

Andrzej Cichocki[a]

[a]Riken, Brain Science Institute, Japan
Laboratory for Advanced Brain Signal Processing
and Warsaw University of Technology, Poland

## ABSTRACT

Blind source separation (BSS) and related methods such as independent component analysis (ICA) and their extensions or sparse component analysis (SCA) refers to wide class of problems in signal and image processing, when one needs to extract the underlying sources from a set of mixture. The goal of BSS can be considered as estimation of true physical sources and parameters of a mixing system, while objective of generalized component analysis (GCA) is finding a new reduced or hierarchical and structured representation for the observed (sensor) multidimensional data that can be interpreted as physically meaningful coding or blind signal decompositions. These methods are generally based on a wide class of unsupervised learning algorithms and they found potential applications in many areas from engineering to neuroscience. The recent trends in blind source separation and generalized component analysis is to consider problems in the framework of matrix factorization or more general signals decomposition with probabilistic generative and tree structured graphical models and exploit some *priori* knowledge about true nature and structure of latent (hidden) components or sources such as spatio-temporal decorrelation, statistical independence, sparsity, nonnegativity, smoothness or lowest possible complexity. The key issue is to find a such transformation or coding which has true physical meaning and interpretation. In this paper we discuss some promising approaches and algorithms for BSS/GCA, especially for ICA and SCA in order to analyze, enhance, perform feature extraction, removing artifacts and denoising of multi-modal, multi-sensory data.

**Keywords:** Independent Component Analysis (ICA) and its extensions. Sparse Component Analysis (SCA), sparse representation, validity and optimality tests

## 1. INTRODUCTION

Data decomposition and representation are widely used in signal processing and neural computing. A traditional method include Fourier analysis and wavelets representations. In many applications is necessary to perform some decomposition of observed signals or data in such way that components have some special properties or structures such as statistical independence, sparsity, smoothness, non-negativity, prescribed statistical distributions and/or specific temporal structure. Recently, several novel methods and approaches have been proposed for decomposition and representations of signals and images, especially, Independent Component Analysis (ICA), Sparse Component Analysis (SCA) and Non-negative Matrix Factorization (NMF).[1–10] All these methods can be expressed algebraically as some specific problems of matrix factorization: Given observation (often called sensor or data) matrix $\mathbf{X} \in \mathbb{R}^{m \times N}$ perform the matrix factorization

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{E}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ represents basis data matrix or mixing matrix (depending on application), $\mathbf{E} \in \mathbb{R}^{m \times N}$ is a matrix representing errors or noise and matrix $\mathbf{S} \in \mathbb{R}^{n \times N}$ contains the corresponding hidden components that give the contribution of each basis vector. Often these components represent unknown source signals with specific temporal structures, features or properties. For example, the rows of matrix $\mathbf{S}$ should be sparse as

Correspondence may be addressed to Dr. A. Cichocki, Riken, Brain Science Institute, Laboratory for Advanced Brain Signal Processing, Wako-shi, Saitama 351-0198 Japan; email sent to *cia@bsp.brain.riken.jp*

Independent Component Analyses, Wavelets, Unsupervised Smart Sensors,
and Neural Networks III, edited by Harold H. Szu, Proceedings of SPIE Vol. 5818
(SPIE, Bellingham, WA, 2005) · 0277-786X/05/$15 · doi: 10.1117/12.606876

possible for SCA or independent as possible for ICA or take only nonnegative values for NMF or values with specific constraints.[6,7] It is important to note that the statistical independence and sparsity are generally different criteria or concepts.

Although some decompositions or matrix factorizations provide an exact reconstruction data (i.e., $\mathbf{X} = \mathbf{AS}$), we shall consider here decompositions which are approximative in nature, however they should be robust to noise and enforce some desirable constraints. In blind source separation (BSS) problem the data matrix $\mathbf{X} = [\mathbf{x}(1), \mathbf{x}(2) \ldots, \mathbf{x}(N)]$ can be represented by vectors $\mathbf{x}(k)$ $(k = 1, 2, \ldots, N)$ for many time instants as multiple measurements or recordings, thus the compact aggregated matrix equation (1) can be written in a vector form as the system of linear equations: $\mathbf{x}(k) = \mathbf{A}\,\mathbf{s}(k) + \mathbf{e}(k)$, where $\mathbf{x}(k) = [x_1(k), \ldots, x_m(k)]^T$ is the vector of the observed signals at the discrete time instant $k$ while $\mathbf{s}(k) = [s_1(k), \ldots, s_n(k)]^T$ is the vector of components at the same time instant. The above formulated problems are related closely to linear inverse problem or more generally, to solving a large ill-conditioned system of linear equations (overdetermined or underdetermined depending on applications) where it is necessary to estimate reliably vectors $\mathbf{s}(k)$ and in some cases also to identify a matrix $\mathbf{A}$ for noisy data.

Different cost functions and imposed constraints may lead to different types of matrix factorizations. For example, in order to find a sparse representation of the matrix $\mathbf{S}$ such that the individual columns of $\mathbf{S}$ should have not only a sparse structure but also a desired sparsity profile and simultaneously some smoothness, we can construct the following optimization problem with two (or more) regularization terms:

$$\min(\frac{1}{2}\|\mathbf{X} - \mathbf{A}\,\mathbf{S}\|_F^2 + \alpha_1 \sum_{k=1}^{N}\sum_{j=1}^{n} |s_j(k)| + \alpha_2 \sum_{k=p}^{N}\sum_{j=1}^{n} |s_j(k) - s_j(k-p)|), \tag{2}$$

where $\alpha_1 \geq 0$ and $\alpha_2 \geq 0$ are regularization coefficients providing sparseness and/or piecewise smoothness. It can be shown that the using optimization approach we can extend this approach to other useful concepts such Smooth Component Analysis (SmoCA) and other spatio-temporal representations with specific features or constraints.

The problems of separating or extracting the original source waveforms from the sensor array, without knowing the transmission channel characteristics and the sources can be expressed briefly as a number of related BSS or blind signal decomposition problems such Independent Component Analysis (ICA) (and its extensions: Topographic ICA, Multidimensional ICA, Kernel ICA, Tree-dependent Component Analysis, Subband Decomposition -ICA), Sparse Component Analysis (SCA), Sparse PCA (SPCA), Non-negative Matrix Factorization (NMF), Smooth Component Analysis (SmoCA), Parallel Factor Analysis (PARAFAC), Time-Frequency Component Analyzer (TFCA) and Multichannel Blind Deconvolution (MBD).[2,4–6,11–14]

## 2. INDEPENDENT COMPONENT ANALYSIS (ICA)AND ITS EXTENSIONS

ICA can be defined as follows: The ICA of a random vector $\mathbf{x}(k) \in \mathbb{R}^m$ is obtained by finding an $n \times m$, (with $m \geq n$), full rank separating (transformation) matrix $\mathbf{W}$ such that the output signal vector $\mathbf{y}(k) = [y_1(k), y_2(k), \ldots, y_n(k)]^T$ (independent components) estimated by

$$\mathbf{y}(k) = \mathbf{W}\,\mathbf{x}(k), \tag{3}$$

are as independent as possible evaluated by an information-theoretic cost function such as minima of Kullback-Leibler divergence.[2,5]

Compared with principal component analysis (PCA), which removes second-order correlations from observed signals, ICA further removes higher-order dependencies. Independence of random variables is a more general concept than decorrelation. Roughly speaking, we say that random variables $y_i$ and $y_j$ are statistically independent if knowledge of the values of $y_i$ provides no information about the values of $y_j$. Mathematically, the independence of $y_i$ and $y_j$ can be expressed by the relationship $p(y_i, y_j) = p(y_i)p(y_j)$, where $p(y)$ denotes the probability density function (pdf) of the random variable $y$. In other words, signals are independent if their joint pdf can be factorized.

If independent signals are zero-mean, then the generalized covariance matrix of $f(y_i)$ and $g(y_j)$, where $f(y)$ and $g(y)$ are different, odd nonlinear activation functions (e.g., $f(y) = \tanh(y)$ and $g(y) = y$ for super-Gaussian sources) is a non-singular diagonal matrix[15]:

$$\mathbf{R}_{\mathbf{f}\,\mathbf{g}} = E\{\mathbf{f}(\mathbf{y})\mathbf{g}^T(\mathbf{y})\} = \begin{bmatrix} E\{f(y_1)g(y_1)\} & & 0 \\ & \ddots & \\ 0 & & E\{f(y_n)g(y_n)\} \end{bmatrix}, \tag{4}$$

i.e., the covariances $E\{f(y_i)g(y_j)\}$ are all zero for $i \neq j$. It should be noted that for odd $f(y)$ and $g(y)$, if the probability density function of each zero-mean source signal is even, then the terms of the form $E\{f(y_i)\}E\{g(y_i)\}$ equal zero. The true general condition for statistical independence of signals is the vanishing of high-order cross-cumulants.[3, 15–17]

The above diagonalization principle can be expressed as[18]

$$\mathbf{R}_{fg}^{-1} = \mathbf{\Lambda}^{-1}, \tag{5}$$

where $\mathbf{\Lambda}$ is any diagonal positive definite matrix (typically, $\mathbf{\Lambda} = \mathbf{I}$ or $\mathbf{\Lambda} = \mathrm{diag}\{\mathbf{R}_{fg}\}$). By pre-multiplying the above equation by separating matrix $\mathbf{W}$ and $\mathbf{\Lambda}$, we obtain:

$$\mathbf{\Lambda}\mathbf{R}_{fg}^{-1}\mathbf{W} = \mathbf{W}, \tag{6}$$

which suggest the following iterative multiplicative learning algorithm

$$\tilde{\mathbf{W}}(l+1) = \mathbf{\Lambda}\mathbf{R}_{fg}^{-1}\mathbf{W}(l), \tag{7}$$

$$\mathbf{W}(l+1) = \tilde{\mathbf{W}}(l+1)\left[\tilde{\mathbf{W}}^T(l+1)\tilde{\mathbf{W}}(l+1)\right]^{-1/2}, \tag{8}$$

where the last equation represents the symmetric orthogonalization to keep algorithm stable. The above algorithm is simple and fast but need prewhitening the data.

In fact, a wide class of ICA algorithms can be expressed in general form as (see Table 1)[4]

$$\nabla\mathbf{W}(l) = \mathbf{W}(l+1) - \mathbf{W}(l) = \eta\mathbf{F}(\mathbf{y})\mathbf{W}(l), \tag{9}$$

where $\mathbf{y}(k) = \mathbf{W}(l)\mathbf{x}(k)$ and the matrix $\mathbf{F}(\mathbf{y})$ can take different forms, for example $\mathbf{F}(\mathbf{y}) = \mathbf{\Lambda}_n - \mathbf{f}(\mathbf{y})\mathbf{g}^T(\mathbf{y})$ with suitably chosen nonlinearities $\mathbf{f}(\mathbf{y}) = [f(y_1), ..., f(y_n)]$ and $\mathbf{g}(\mathbf{y}) = [g(y_1), ..., g(y_n)]$.[4, 16, 19–21]

Assuming prior knowledge of the source distributions $p_i(y_i)$, we can estimate $\mathbf{W}$ using maximum likelihood (ML):

$$J(\mathbf{W}, \mathbf{y}) = -\frac{1}{2}\log|\det(\mathbf{W}\mathbf{W}^T)| - \sum_{i=1}^{n}\log(p_i(y_i) \tag{10}$$

Using natural gradient descent to increase likelihood we get:

$$\mathbf{W}(l+1) = \eta\ \left[\mathbf{I} - \mathbf{f}(\mathbf{y})\mathbf{y}^T\right]\mathbf{W}(l), \tag{11}$$

where $\mathbf{f}(\mathbf{y}) = [f_1(y_1), f_2(y_2), \ldots, f_n(y_n)]^T$ is an entry-wise nonlinear score function defined by

$$f_i(y_i) = -\frac{p_i'(y_i)}{p_i(y_i)} = -\frac{d\log(p_i(y_i)}{d(y_i)}. \tag{12}$$

It should be noted that ICA can perform blind source separation, i.e., enable to estimate true sources only if they are all statistically independent and non Gaussian (except possibly of one).[4, 22]

**Table 1.** Basic equivariant adaptive learning algorithms for ICA. Some of these algorithms require prewhitening.

| No. | Learning Algorithm | References |
|---|---|---|
| 1. | $\Delta \mathbf{W} = \eta \Big[\boldsymbol{\Lambda} - \langle \mathbf{f}(\mathbf{y})\, \mathbf{g}^T(\mathbf{y})\rangle\Big] \mathbf{W}$ | Cichocki, Unbehauen, Rummert (1994) |
| | $\boldsymbol{\Lambda}$ is a diagonal matrix with nonnegative elements $\lambda_{ii}$ | |
| | $\mathbf{W}(l+1) = \Big[\mathbf{I} \mp \eta\, [\mathbf{I} - \langle \mathbf{f}(\mathbf{y})\, \mathbf{g}^T(\mathbf{y})\rangle]\Big]^{\mp 1} \mathbf{W}(l)$ | Cruces, Cichocki, Castedo (2000) |
| 2. | $\Delta \mathbf{W} = \eta \Big[\boldsymbol{\Lambda} - \langle \mathbf{f}(\mathbf{y})\, \mathbf{y}^T\rangle\Big] \mathbf{W}, \quad f(y_i) = -p'(y_i)/p(y_i)$ | Amari, Cichocki, Yang (1995) |
| | $\lambda_{ii} = \langle f(y_i(k)) y_i(k)\rangle$ or $\lambda_{ii} = 1, \ \forall i$ | Bell, Sejnowski (1995) |
| | | Amari, Chen, Cichocki (1999) |
| 3. | $\Delta \mathbf{W} = \eta \Big[\mathbf{I} - \langle \mathbf{y}\, \mathbf{y}^T\rangle - \langle \mathbf{f}(\mathbf{y})\, \mathbf{y}^T\rangle + \langle \mathbf{y}\, \mathbf{f}^T(\mathbf{y})\rangle\Big] \mathbf{W}$ | Cardoso, Laheld, (1996) |
| 4. | $\Delta \mathbf{W} = \eta \Big[\mathbf{I} - \langle \mathbf{y}\, \mathbf{y}^T\rangle - \langle \mathbf{f}(\mathbf{y})\, \mathbf{y}^T\rangle + \langle \mathbf{f}(\mathbf{y})\, \mathbf{f}^T(\mathbf{y})\rangle\Big] \mathbf{W}$ | Karhunen, Pajunen (1997) |
| 5. | $\tilde{\mathbf{W}} = \mathbf{W} + \eta \Big[\boldsymbol{\Lambda} - \langle \mathbf{f}(\mathbf{y})\, \mathbf{y}^T\rangle\Big] \mathbf{W}, \ \ \lambda_{ii} = \langle f(y_i)\ y_i\rangle$<br>$\eta_{ii} = [\lambda_{ii} + \langle f'(y_i)\rangle]^{-1}; \quad \mathbf{W} = \tilde{\mathbf{W}}(\tilde{\mathbf{W}}^T \tilde{\mathbf{W}})^{-1/2}$ | Hyvärinen, Oja (1999) |
| 6. | $\tilde{\mathbf{W}} = \boldsymbol{\Lambda}\, \mathbf{R}_{fg}^{-1} \mathbf{W}$ | Fiori (2003) |
| | $\mathbf{W} = \tilde{\mathbf{W}}(\tilde{\mathbf{W}}^T \tilde{\mathbf{W}})^{-1/2}$ | |
| 7. | $\Delta \mathbf{W} = \eta \Big[\mathbf{I} - \boldsymbol{\Lambda}^{-1} \langle \mathbf{y}\, \mathbf{y}^T\rangle\Big] \mathbf{W}$ | Amari, Cichocki (1998) |
| | $\lambda_{ii}(k) = \langle y_i^2(k))\rangle$ | Choi, Cichocki, Amari (2000) |
| 8. | $\Delta \mathbf{W} = \eta \Big[\mathbf{I} - \mathbf{C}_{1,q}(\mathbf{y}, \mathbf{y})\, \mathbf{S}_{q+1}(\mathbf{y})\Big] \mathbf{W}$<br>$C_{1,q}(y_i, y_j) = Cum(y_i, \underbrace{y_j, \ldots, y_j}_{q})$ | Cruces, Castedo, Cichocki (2002) |
| 9. | $\mathbf{W}(l+1) = \exp(\eta\, \mathbf{F}[\mathbf{y}])\, \mathbf{W}(l)$ | Nishimori, Fiori(1999,2003) |
| | $\mathbf{F}(\mathbf{y}) = \boldsymbol{\Lambda} - \langle \mathbf{y}\, \mathbf{y}^T\rangle - \langle \mathbf{f}(\mathbf{y})\, \mathbf{y}^T\rangle + \langle \mathbf{y}\, \mathbf{f}^T(\mathbf{y})\rangle$ | Cichocki, Georgiev (2002) |
| 10. | $\Delta \mathbf{W} = \eta \mathbf{F}[\mathbf{y}] \mathbf{W}$ | Amari, (1997) |
| | $f_{ij} = \left[\lambda_{ii}\delta_{ij} - \alpha_{1i} \left\langle y_i y_j^* \right\rangle - \alpha_{2i} < f(y_i) y_j^* >\right]$ | Amari *et al.* (2000) |

### 2.1. Sequential Blind Source Extraction

There are two main approaches to solve the problem of blind separation and deconvolution. The first approach, which was mentioned briefly in the previous section, is to simultaneously decompose or separate all sources. In the second one, we extract sources sequentially in a blind fashion, one by one, rather than separating them all simultaneously. In many applications, a large number of sensors (electrodes, sensors, microphones or transducers) are available but only a very few source signals are subjects of interest. For example, in the modern EEG or MEG devices, we observe typically more than 100 sensor signals, but only a few source signals are interesting; the rest can be considered as interfering noise. In another example, the cocktail party problem, it is usually essential to extract the voices of specific persons rather than separate all the source signals of all speakers available (in mixing form) from an array of microphones. For such applications it is essential to develop and apply reliable, robust and effective learning algorithms which enable us to extract only a small number of source signals that are potentially interesting and contain useful information.

We can use two different models and criteria. The first criterion is based on higher order statistics (HOS) which assumes that the sources are mutually statistically independent and non-Gaussian (at most only one can be Gaussian). For independence criteria, we will use some measures of non-Gaussianity.[4] Let us assume that observed (sensor) signals are prewhitened (sphered), and as a cost function for minimization, we may employ[4, 5]

$$\mathcal{J}(\mathbf{w}_1) = -\frac{1}{4}\left|\kappa_4(y_1)\right| = -\frac{\beta}{4}\kappa_4(y_1), \tag{13}$$

where $\kappa_4(y_1)$ is the normalized kurtosis defined for zero-mean signals by $\kappa_4(y_1) = E\{|y_1|^2\} - 3$ and the parameter $\beta$ determines the sign of the kurtosis of the extracted signal. The gradient of the above cost function can be expressed as $\nabla_{\mathbf{w}_1}\mathcal{J} = -\beta[E\{y_1^3\mathbf{x}\} - 3E\{y_1^2\}E\{y_1\mathbf{x}\}]$. This leads to modified and improved fast ICA algorithm expressed as

$$\begin{aligned}
\tilde{\mathbf{w}}_1(l+1) &= -\beta[E\{y_1^3\mathbf{x}\} - 3E\{y_1^2\}E\{y_1\mathbf{x}\}], \quad y_1 = \mathbf{w}_1^T(l)\mathbf{x}, & (14)\\
\mathbf{w}_1(l+1) &= \frac{\tilde{\mathbf{w}}_1(l+1)}{||\tilde{\mathbf{w}}_1(l+1)||_2}. & (15)
\end{aligned}$$

The above algorithm is more robust in respect to the number of samples in comparison to original Fast ICA algorithm.[5]

The second alternative criterion, based on the concept of linear predictability and assumes that source signals have some temporal structure, i.e., the sources are colored with different autocorrelation functions or equivalently have different spectra shapes. In this approach, we exploit the temporal structure of signals rather than their statistical independence.[23, 24] Intuitively speaking, the source signals $s_j$ have less complexity than the mixed sensor signals $x_j$. In other words, the degree of temporal predictability of any source signal is higher than (or equal to) that of any mixture.

For example, waveforms of a mixture of two sine waves with different frequencies are more complex or less predictable than either of the original sine waves. This means that applying the standard linear predictor model and minimizing the mean squared error $E\{\varepsilon^2\}$, which is measure of predictability, we can separate or extract signals with different temporal structures. More precisely, by minimizing the error, we maximize a measure of temporal predictability for each recovered signal.[25, 26]

It is worth to note that two criteria used in BSE: temporal linear predictability and non-Gaussianity based on kurtosis may lead to different results. Temporal predictability forces the extracted signal to be smooth and possibly less complex while the non-Gaussianity measure forces the extracted signals to be as independent as possible with sparse representation for sources that have positive kurtosis.

Let us assume for simplicity, that we want to extract only one source signal, e.g. $s_j(k)$, from the available sensor vector $\mathbf{x}(k)$. For this purpose, we employ a single processing unit described as (see Figure 1):

$$y_1\left(k\right) \;=\; \mathbf{w}_1^T\mathbf{x}\left(k\right) = \sum_{i=1}^{m} w_{1i}\, x_i\left(k\right), \tag{16}$$

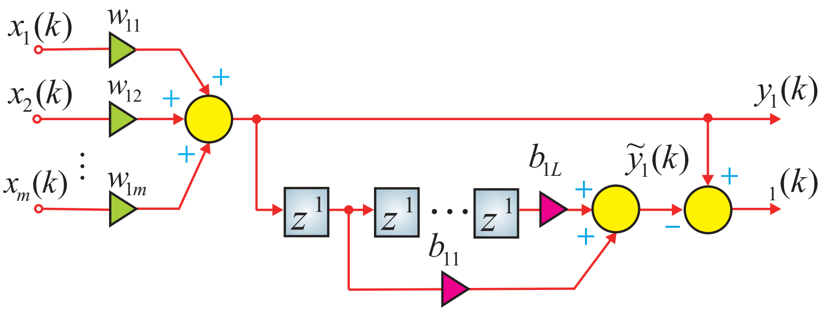

**Figure 1.** The neural network structure of single extraction unit using a linear predictor.

$$\varepsilon_1(k) \quad = \quad y_1(k) - \sum_{p=1}^{L} b_{1p}\, y_1(k-p) = \mathbf{w}_1^T \mathbf{x}(k) - \mathbf{b}_1^T \bar{\mathbf{y}}_1(k), \tag{17}$$

where $\mathbf{w}_1 = [w_{11}, w_{12}, \ldots, w_{1m}]^T$, $\bar{\mathbf{y}}_1(k) = [y_1(k-1), y_1(k-2), \ldots, y_1(k-L)]^T$, $\mathbf{b}_1 = [b_{11}, b_{12}, \ldots, b_{1L}]^T$ and $B_1(z) = \sum\limits_{p=1}^{L} b_{1p} z^{-p}$ is the transfer function of the corresponding FIR filter. It should be noted that the FIR filter can have a sparse representation. In particular, only one single processing unit, e.g. with delay $p$ and $b_{1p} \neq 0$ can be used instead of $L$ parameters. The processing unit has two outputs: $y_1(k)$ which estimates the extracted source signals, and $\varepsilon_1(k)$, which represents a linear prediction error or estimation of the innovation, after passing the output signal $y_1(k)$ through FIR filter.

Our objective is to estimate optimal values of vectors $\mathbf{w}_1$ and $\mathbf{b}_1$, in such a way that the processing unit successfully extracts one of the sources. This is achieved if the global vector defined as $\mathbf{g}_1 = \mathbf{A}^T \mathbf{w}_1 = \left(\mathbf{w}_1^T \mathbf{A}\right)^T = c_j \mathbf{e}_j$ contains only one nonzero element, e.g. in the $j$-th row, such that $y_1(k) = c_j s_j$, where $c_j$ is an arbitrary nonzero scaling factor. For this purpose, we reformulate the problem as a minimization of the cost function

$$\mathcal{J}(\mathbf{w}_1, \mathbf{b}_1) = E\left\{\varepsilon_1^2\right\}. \tag{18}$$

The main motivation for applying such a cost function is the assumption that primary source signals (signals of interest) have temporal structures and can be modelled, e.g., by an autoregressive model.[4, 27, 28]

According to the AR model of source signals, the filter output can be represented as $\varepsilon_1(k) = y_1(k) - \tilde{y}_1(k)$, where $\tilde{y}_1(k) = \sum_{p=1}^{L} b_{1p} y_1(k-p)$ is defined as an error or estimator of the innovation source $\tilde{s}_j(k)$. The mean squared error $E\{\varepsilon_1^2(k)\}$ achieves a minimum $c_1^2 E\{\tilde{s}_j^2(k)\}$, where $c_1$ is a positive scaling constant, if and only if $y_1 = \pm c_1 s_j$ for any $j \in \{1, 2, \ldots, m\}$ or $y_1 = 0$ holds.

Let us consider the processing unit shown in Figure 1. The associated cost function (18) can be evaluated as follows:

$$E\left\{\varepsilon_1^2\right\} = \mathbf{w}_1^T \widehat{\mathbf{R}}_{\mathbf{x}_1\mathbf{x}_1} \mathbf{w}_1 - 2\mathbf{w}_1^T \widehat{\mathbf{R}}_{\mathbf{x}_1\bar{\mathbf{y}}_1} \mathbf{b}_1 + \mathbf{b}_1^T \widehat{\mathbf{R}}_{\bar{\mathbf{y}}_1\bar{\mathbf{y}}_1} \mathbf{b}_1, \tag{19}$$

where $\widehat{\mathbf{R}}_{\mathbf{x}_1\mathbf{x}_1} \approx E\{\mathbf{x}_1\mathbf{x}_1^T\}$, $\widehat{\mathbf{R}}_{\mathbf{x}_1\bar{\mathbf{y}}_1} \approx E\{\mathbf{x}_1\bar{\mathbf{y}}_1^T\}$ and $\widehat{\mathbf{R}}_{\bar{\mathbf{y}}_1\bar{\mathbf{y}}_1} \approx E\{\bar{\mathbf{y}}_1\bar{\mathbf{y}}_1^T\}$, are estimators of true values of correlation and cross-correlation matrices: $\mathbf{R}_{\mathbf{x}_1\mathbf{x}_1}, \mathbf{R}_{\mathbf{x}_1\bar{\mathbf{y}}_1}, \mathbf{R}_{\bar{\mathbf{y}}_1\bar{\mathbf{y}}_1}$, respectively. In order to estimate vectors $\mathbf{w}_1$ and $\mathbf{b}_1$, we evaluate gradients of the cost function and equalize them to zero as follows:

$$\frac{\partial \mathcal{J}_1(\mathbf{w}_1, \mathbf{b}_1)}{\partial \mathbf{w}_1} \quad = \quad 2\widehat{\mathbf{R}}_{\mathbf{x}_1\mathbf{x}_1} \mathbf{w}_1 - 2\widehat{\mathbf{R}}_{\mathbf{x}_1\bar{\mathbf{y}}_1} \mathbf{b}_1 = \mathbf{0}, \tag{20}$$

$$\frac{\partial \mathcal{J}_1(\mathbf{w}_1, \mathbf{b}_1)}{\partial \mathbf{b}_1} \quad = \quad 2\widehat{\mathbf{R}}_{\bar{\mathbf{y}}_1\bar{\mathbf{y}}_1} \mathbf{b}_1 - 2\widehat{\mathbf{R}}_{\bar{\mathbf{y}}_1\mathbf{x}_1} \mathbf{w}_1 = \mathbf{0}. \tag{21}$$

Solving the above matrix equations and assume that for pre-whitened data $\widehat{\mathbf{R}}_{\mathbf{x}_1\mathbf{x}_1} = \mathbf{I}$, we obtain a simple iterative algorithm:

$$\begin{aligned} \tilde{\mathbf{w}}_1 &= \widehat{\mathbf{R}}_{\mathbf{x}_1\bar{\mathbf{y}}_1}\mathbf{b}_1, \;\; \mathbf{w}_1 = \frac{\tilde{\mathbf{w}}_1}{||\tilde{\mathbf{w}}_1||_2}, \qquad (22) \\ \mathbf{b}_1 &= \widehat{\mathbf{R}}^{-1}_{\bar{\mathbf{y}}_1\bar{\mathbf{y}}_1}\widehat{\mathbf{R}}_{\bar{\mathbf{y}}_1\mathbf{x}_1}\mathbf{w}_1 = \widehat{\mathbf{R}}^{-1}_{\bar{\mathbf{y}}_1\,\bar{\mathbf{y}}_1}\widehat{\mathbf{R}}_{\bar{\mathbf{y}}_1\,y_1}, \qquad (23) \end{aligned}$$

where the matrices $\widehat{\mathbf{R}}_{\bar{\mathbf{y}}_1\,\bar{\mathbf{y}}_1}$ and $\widehat{\mathbf{R}}_{\bar{\mathbf{y}}_1\,y_1}$ are estimated based on the parameters $\mathbf{w}_1$ obtained in the previous iteration step. In order to avoid the trivial solution $\mathbf{w}_1 = \mathbf{0}$, we normalize the vector $\mathbf{w}_1$ to unit length in each iteration step as $\mathbf{w}_1(l+1) = \tilde{\mathbf{w}}_1(l+1)/\left\|\tilde{\mathbf{w}}_1(l+1)\right\|_2$ (which ensures that $E\{y_1^2\} = 1$).

It is worth to note here that in our derivation matrices $\widehat{\mathbf{R}}_{\bar{\mathbf{y}}_1\,\bar{\mathbf{y}}_1}$ and $\widehat{\mathbf{R}}_{\bar{\mathbf{y}}_1\,y_1}$ are assumed to be independent of the vector $\mathbf{w}_1(l+1)$, i.e., they are estimated based on $\mathbf{w}_1(l)$ in the previous iteration step. This two-phase procedure is similar to the expectation maximization (EM) scheme: (i) Freeze the correlation and cross-correlation matrices and learn the parameters of the processing unit $(\mathbf{w}_1, \mathbf{b}_1)$; (ii) freeze $\mathbf{w}_1$ and $\mathbf{b}_1$ and learn new statistics (i.e., matrices $\widehat{\mathbf{R}}_{\bar{\mathbf{y}}_1\,y_1}$ and $\mathbf{R}_{\bar{\mathbf{y}}_1\,\bar{\mathbf{y}}_1}$) of the estimated source signal, then go back to (i) and repeat. Hence, in phase (i), our algorithm extracts a source signal, whereas in phase (ii) it learns the statistics of the source.

The derived algorithm is similar to the power method for finding the eigenvector $\mathbf{w}_1$ associated with the maximal eigenvalue of the matrix $\mathbf{R}_{\mathbf{x}_1}(\mathbf{b}_1) = E\{\sum_{p=1}^{L} b_{1p}\mathbf{x}_1(k)\mathbf{x}_1^T(k-p)\}$. This observation suggests that it is not necessary to minimize the cost function with respect to parameters $\{b_{1p}\}$ but it is enough to choose an arbitrary set of them for which the largest eigenvalue is unique (single). More generally, if all eigenvalues of the generalized covariance matrix $\mathbf{R}_{\mathbf{x}_1}(\mathbf{b}_1)$ are distinct, then we can extract all sources simultaneously by estimating principal eigenvectors of $\mathbf{R}_{\mathbf{x}_1}(\mathbf{b}_1)$.

For noisy data, instead of linear predictor, we can use a bandpass filter (or in a parallel way several processing units with a bank of bandpass filters) with fixed or adjustable center frequency and a bandpass bandwidth.[26, 29, 30]

### 2.2. Multiresolution Subband Decomposition – Independent Component Analysis (MSD-ICA)

Despite the success of using standard ICA in many applications, the basic assumptions of ICA may not hold for some kind of signals hence some caution should be taken when using standard ICA to analyze real world problems, especially in analysis of EEG/MEG data. In fact, by definition, the standard ICA algorithms are not able to estimate statistically dependent primary sources, that is, when the independence assumption is violated (even for very weak dependence). In this section, we will present a natural extension and generalization of ICA called Multiresolution Subband Decomposition ICA (MSD-ICA) which relaxes considerably the assumption regarding mutual independence of primarily sources.[4, 31–33] The key idea in this approach is the assumption that the unknown wide-band source signals can be dependent, however some their subcomponents are independent. In other words, we assume that each unknown source can be modelled or represented in the time domain or a transform domain as a sum (or linear combinations) of sub-components (were some of them are hopefully independent):

$$s_i(k) = s_{i1}(k) + s_{i2}(k) + \cdots + s_{iK}(k). \qquad (24)$$

For example, in the simplest case, source signals can be modelled or decomposed into their low- and high-frequency sub-components: $s_i(k) = s_{iL}(k) + s_{iH}(k) \quad (i = 1, 2, \ldots, n)$. In practice, the high-frequency sub-components $s_{iH}(k)$ are often found to be mutually independent, while the low-frequency sub-components are weakly dependent. In such a case, we can use a High Pass Filter (HPF) to extract mixture of the high frequency sub-components and then apply any standard ICA algorithm to such preprocessed sensor (observed) signals.

Let us assume that only a certain set of sub-components are independent. Provided that for some of the frequency subbands (at least one) all sub-components, say $\{s_{ij}(k)\}_{i=1}^{n}$, are mutually independent or temporally decorrelated, then we can easily estimate the mixing or separating system under condition that these subbands can be identified by some *a priori* knowledge or detected by some self-adaptive process. For this purpose, we

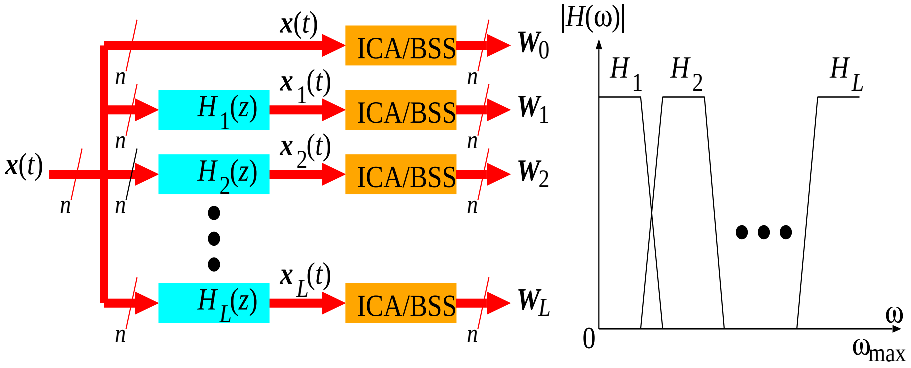

**Figure 2.** Bank of filters employed in preprocessing stage for MSD-ICA with typical frequency bands. For each sensor signal we employ the identical set of filters. The sub-bands can be overlapped or not and have more complex sub-bands forms. The basic concept in Subband Decomposition ICA is to divide the sensor signal spectra into their subspectra or subbands, and then to treat those subspectra individually for the purpose at hand. The subband signals can be ranked and processed independently.

simply apply any standard ICA algorithm, however not only for all available raw sensor data but also for suitably pre-processed or decomposed (e.g., subband filtered) sensor signals (see Fig. 2).

By applying any standard ICA/BSS algorithm for specific sub-bands and raw sensor data, we obtain sequence of separating matrices $\mathbf{W}_0, \mathbf{W}_1, \ldots, \mathbf{W}_L$, where $\mathbf{W}_0$ is the separating matrix from the original data $\mathbf{x}$ and $\mathbf{W}_j$ is the separating matrix from preprocessing sensor data $\mathbf{x}_j$ in the $j$-th sub-band. In order to estimate true separating and mixing matrices and to identify for which sub-bands corresponding source subcomponents are independent, we propose to compute the global (mixing-separating) matrices $\mathbf{G}^{pq} = \mathbf{W}_p \mathbf{W}_q^{-1}$, $\forall p \neq q$ and $m = n$, where $\mathbf{W}_q$ is estimating separating matrix for $q$-th sub-band. If subcomponents are mutually independent for at least two sub-bands, say for the sub-band $p$ and sub-band $q$, then the global matrix $\mathbf{W}_p \mathbf{W}_q^{-1} = \mathbf{P}^{pq}$ will be generalized permutation matrix with only one nonzero (or dominated) element in each row and each column. This follows from the simple observation that in such case the both matrices $\mathbf{W}_p$ and $\mathbf{W}_q$ represent inverses (for $m = n$) of the same mixing matrix $\mathbf{A}$ (neglecting nonessential scaling and permutation ambiguities). In this way, we can blindly identify essential information for which frequency sub-bands the source subcomponents are independent and we can easily identify correctly the mixing matrix. Furthermore, the same concept can be used to estimate blindly the performance index and to compare performance of various ICA algorithms, especially for large scale problems.

In the preprocessing stage we can use any linear transforms, especially, more sophisticated decomposition methods, such as block transforms, multirate sub-band filter bank or wavelet transforms, can be applied. We can extend and generalize further this concept by performing the decomposition of sensor signals in a composite time-frequency domain rather than in frequency sub-bands as such. This naturally leads to the concept of wavelets packets (sub-band hierarchical trees) and to block transform packets.[4, 10, 11] Such preprocessing techniques has been extensively tested and implemented in ICALAB.[34]

Such explanation can be summarized as follows. The MSD-ICA (Multiresolution Subband Decomposition ICA) can be formulated as a task of estimation of the separating matrix $\mathbf{W}$ and/or the estimating mixing matrix $\hat{\mathbf{A}}$ on the basis of suitable wavelet package or subband decomposition of sensor signals and by applying a classical ICA (instead for raw sensor data) for one or several preselected subbands for which source sub-components are independent.

### 2.3. Validity of ICA, BSS algorithms for real world data

One of the fundamental question in BSS is problem whether the obtained results of the specific BSS/ICA algorithm is reliable and represent inherent properties of the model and data or it is just a random, synthetic or purely mathematical, decomposition of data without any physical meaning. In fact, since most of BSS algorithms are stochastic in nature, their results could be somewhat different in different runs even for the same algorithm. Thus, the results obtained in a single run or for single set of data of any BSS algorithm should be interpreted with reserve and reliability of estimated sources should be analyzed by investigating the spread of the obtained estimates for many runs.[35] Such an analysis can be performed, for example by using resampling or bootstrapping method in which the available data is randomly changed by producing surrogate data sets from the original data.[36] The specific ICA/BSS algorithm is then run many times with bootstrapped samples that are somewhat different from each other. Alternative approach called ICASSO which is based on running the specific BSS algorithm many times for various different initial conditions and parameters and by visualizing the clustering structure of the estimated sources (components) in the signal subspace.[35]

It is worth to note that the concept of MSD-ICA described in the previous section can be extended to more general and flexible multi-dimensional models for checking validity and reliability of ICA (or more generally BSS) algorithms (see Figure 2). In this model we can use a bank of stable filters with transfer functions $H_i(z)$, for example, set of FIR (finite impulse response filters). The parameters (coefficients) of such FIR filters can be suitably designed or even be randomly generated. In this case, the proposed method has some similarity with resampling or bootstrap approach.[36] Similarly as in MSD-ICA, we can run any BSS algorithm for sufficiently large number $L$ of filters and generate set of separating matrices: $\{\mathbf{W}_0, \mathbf{W}_1, \ldots, \mathbf{W}_L\}$ or alternatively set of estimated mixing matrices: $\{\hat{\mathbf{A}}_0, \hat{\mathbf{A}}_1, \ldots, \hat{\mathbf{A}}_L\}$. In the next step we estimate the set of global mixing-separating matrices $\mathbf{G}^{pq} = \mathbf{W}_p \mathbf{W}_q^+$ for any $p \neq q$.

The performance of blind separation can be characterized by one single performance index (sometimes referred as Amari's performance index) which we refer as blind performance index (since we do not know a true mixing matrix):

$$BPI_i^{pq} = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{\sum_{i=1}^{n}|g_{ij}^{pq}|^2}{max_i\ |g_{ij}^{pq}|^2} - 1\right) + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{n}|g_{ij}^{pq}|^2}{max_j\ |g_{ij}^{pq}|^2} - 1\right), \tag{25}$$

where $g_{ij}^{pq}$ is $ij$-th element of the matrix $\mathbf{G}^{pq}$. In many cases, we are not able to achieve perfect separation for some sources or we are able to extract only some sources (not of all them). In such cases instead of using one global performance index, we can define local performance index as

$$BPI_i^{pq} = \left(\frac{\sum_{j=1}^{n}|g_{ij}^{pq}|^2}{max_j\ |g_{ij}^{pq}|^2} - 1\right). \tag{26}$$

If the performance index $BPI_i^{pq}$ for specific index $i$ and filters $p, q$ is close to zero this means that with high probability this component is successfully extracted. In order to asses significant components the all estimated components should be clustered according their mutual similarities. These similarities can be searched in the time domain or the frequency domain. The natural measure of similarity between of the estimated components can be absolute value of their mutual correlation coefficients $|r_{ij}|$ for $i \neq j$ which are elements of the similarity matrix[35]

$$\mathbf{R} = \overline{\mathbf{W}}\, \mathbf{R}_{xx}\, \overline{\mathbf{W}}^T, \tag{27}$$

where $\overline{\mathbf{W}} = [\mathbf{W}_0, \mathbf{W}_1, \ldots, \mathbf{W}_L]$ and $\mathbf{R}_{xx} = E\{\mathbf{x}\mathbf{x}^T\} = \mathbf{A}\mathbf{R}_{ss}\mathbf{A}^T$ is covariance matrix of observations under assumption that the covariance matrix of sources $\mathbf{R}_{ss} = E\{\mathbf{s}\mathbf{s}^T\}$ is a diagonal matrix and separating matrices $\mathbf{W}_p$ are normalized (e.g., to unit length vectors).

In summary, in order to estimate algorithmic reliability it is necessary to run the BSS/ICA algorithm many times using different initial conditions and decomposed (preprocessed data and assessing which of the basis vectors and corresponding components are found in almost all run. For this purpose the estimated components should be clustered and classified. The reliable components corresponds to small and well separated clusters from the rest of components, while unreliable components usually do not belong to any cluster.

## 3. SPARSE COMPONENT ANALYSIS AND SPARSE SIGNAL REPRESENTATIONS

Sparse Component Analysis (SCA) and sparse signals representations (SSR) arise in many scientific problems, especially, where we wish to represent signals of interest by using a small (or sparse) number of basis signals from a much larger set of signals, often called dictionary.[7] Such problems arise also in many applications such as electro-magnetic and biomagnetic inverse problems (EEG/MEG), feature extraction, filtering, wavelet denoising, time-frequency representation, neural and speech coding, spectral estimation, direction of arrival estimation, failure diagnosis and speed-up processing.[4, 10, 37–39]

In opposite to ICA where the mixing matrix and source signals are estimated simultaneously the SCA is usually a multi stage procedure. In first stage we need to find a suitable linear transformation which guarantee that sources in the transformed domain are sufficiently sparse. Typically, we represent the observed data in the time-frequency domain using wavelets package.[38] In the next step, we estimate the columns $\mathbf{a}_i$ of the mixing matrix $\mathbf{A}$ using a sophisticated hierarchical clustering technique. This step is the most difficult and challenging task since it requires to identify precisely intersections of all hyperplanes on which observed data are located.[37, 40] In the last step, we estimate sparse sources using for example a modified robust linear programming (LP), quadratic programming (QP) or semi-definite programming (SDP) optimization. The big advantage of SCA is its ability to reconstruct of original sources and also their numbers even if the number of observations (sensors) is smaller than number of sources under certain weak conditions.[33, 38, 39] Moreover, the system can be highly nonstationary (i.e., the number of active sources can change dramatically in time) and sources can be statistically dependent.

In fact, finding a sparse (or in many cases the sparsest) solution can be viewed equivalently as the problem of selecting very few columns $\mathbf{a}_j \in \mathbb{R}^m$ (called atoms) of the matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n]$ (called dictionary) to represent the observation vector $\mathbf{x}$ which is referred as the subset selection problem. In time-frequency and wavelets theory communities this process is commonly referred to as atomic decomposition, since we decompose the signals $\mathbf{x}$ into their building atoms $\mathbf{a}_j$, taken from the dictionary $\mathbf{A}$. We can state the subset selection problem as follows: Find an optimal subset of $r << n$ columns from the matrix $\mathbf{A}$, which we denote by $\mathbf{A}_r \in \mathbb{R}^{m \times r}$ such that $\mathbf{A}_r \mathbf{s}_{r*} \cong \mathbf{x}$, or equivalently $\mathbf{A}_r \mathbf{s}_{r*} + \mathbf{e}_r = \mathbf{x}$, where $\mathbf{e}_r$ represents some residual error vector which norm should be as small as possible or at below some threshold. Usually, we have interest in sparsest and unique representation, i.e., it is necessary to find solution having the smallest possible number of non-zero-components. The problem can be reformulated as the following optimization problem:[33]:

$$(P_\rho) \quad J_\rho(\mathbf{s}) = \|\mathbf{s}\|_\rho = \sum_{j=1}^{n} \rho(s_j) \quad \text{s. t.} \qquad \mathbf{A}\,\mathbf{s} = \mathbf{x}, \tag{28}$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$, (usually with $n >> m$) and $\|\mathbf{s}\|_\rho$ suitably chosen function which measures the sparsity of the vector $\mathbf{s}$. It should be noted the sparsity measure does not need be necessary a norm, although we use such notation. For example, we can apply Shannon, Gauss or Renyi entropy or normalized kurtosis as measure of the sparsity.[4, 10, 41] In the standard form, we use $l_p$-norm ($\|\mathbf{s}\|_p$) with $0 \le p \le 1$. Especially, $l_0$ quasi-norm attract a lot of attention since it ensures sparsest representation.[7, 37, 38] Unfortunately, such formulated problem (28) for $l_p$-norm with $p < 1$ is very difficult to solve, especially for $p = 0$ it is NP-hard, so for a large scale problem it is numerically untractable. For this reason, we often use Basis Pursuit (BP) which employs a standard Linear Programming (LP) for $\|\mathbf{s}\|_\rho = \|\mathbf{s}\|_1$ subject to constraints $\mathbf{A}\mathbf{s} = \mathbf{x}$.

In practice, due to noise and other uncertainty (e.g., measurement errors) the system of linear underdetermined equations should not be satisfied precisely but with some prescribed tolerance (i.e., $\mathbf{A}\,\mathbf{s} \cong \mathbf{x}$ in the sense that $\|\mathbf{x} - \mathbf{A}\,\mathbf{s}\|_q \le \varepsilon$). From the practical point of view as well as from a statistical point of view, it is convenient and quite natural to replace the exact constraints $\mathbf{x} = \mathbf{A}\,\mathbf{s}$ by the constraint $\|\mathbf{x} - \mathbf{A}\,\mathbf{s}\|_q \le \varepsilon$, where choice of $l_q$-norm depends on distribution of noise and specific applications. For noisy and uncertain data we should to use a more flexible and robust cost function (in comparison to the standard $(P_\rho)$ problem) which will be referred as Extended Basis Pursuit Denoising $(EBPD)$[33]:

$$(EBPD) \qquad J_{q,\rho}(\mathbf{s}) = \|\mathbf{x} - \mathbf{A}\,\mathbf{s}\|_q^q + \alpha\, \|\mathbf{s}\|_\rho, \tag{29}$$

which can be often solved using standard quadratic programming (QP) or semi-definite programming (SDP).

There are several possible basic choices for $l_q$ and sparsity criteria ($\|\mathbf{s}\|_\rho = \|\mathbf{s}\|_p$) For example, for the uniform (Laplacian) distributed noise we should choose $l_\infty$-Chebyshev norm ($l_1$-norm). Some basic choices of $\rho$ (for $l_q = 2$) are $\rho = 0$ (minimum $l_0$ quasi norm or atomic decomposition related with the matching pursuit (MP) and FOCUSS algorithm), $\rho = 1$ (basis pursuit denoising) and $\rho = 2$ (ridge regression).[7, 10, 41] The optimal choice of $\rho$ norms depends on distribution of noise in sparse components. For example, for noisy components, we can use robust norms such as Huber function defined as $\|\mathbf{s}\|_{\rho_H} = \sum_i \rho_H(s_i)$, where $\rho_H(s_i) = s_i^2/2$ if $|s_i| \leq \beta$ and $\rho_H(s_i) = \beta\,|s_i| - \beta^2/2$ if $|s_i| > \beta$, and/or epsilon norm defined as $\|\mathbf{s}\|_\varepsilon = \sum_j |s_j|_\varepsilon$ where $|s_j|_\varepsilon = \max\{0,\ (|s_j| - \varepsilon)\}$.

The practical importance of the $EBPD$ approach in comparison to the standard LP or BP approach is that the $EPBD$ allows for treating the presence of noise or errors due to mismodeling. Moreover, using the $EBPD$ approach, we have possibility to adjust the sparsity profile (i.e., adjust the number of non-zero components) by tuning the parameter $\alpha$. In contrast, by using the LP approach we do not have such option. Furthermore, the method can be applied both for undercomplete and/or overcomplete models (i.e., when the number of sources is larger or less than the number of sensors. It should be noted that if the regularization parameter $\alpha$ is too large or too small the optimal solution of the $EBPD$ can be useless. By increasing $\alpha$, we increase sparsity of $\mathbf{s}_*$ till all entries will be zero for $\alpha \geq \|\mathbf{A}^T\mathbf{x}\|_\rho^*$ (where $\|\cdot\|_\rho^*$ denotes dual norm to $\|\cdot\|_\rho$). More precisely, for $0 < \alpha \leq \|\mathbf{A}^T\mathbf{x}\|_\rho^*$ the solution $\mathbf{s}_*$ are a piecewise linear function of $\alpha$, under condition that the solution is unique. In the general case, however, the number of non-zero elements of $\mathbf{s}_*$ is not necessarily a monotonic function of $\alpha$. In the special case, when basis matrix $\mathbf{A}$ is orthogonal, i.e., $\mathbf{A}\mathbf{A}^T = \mathbf{A}^T\mathbf{A} = \mathbf{I}_n$ with $m = n$ the $EQP$ problem for $q = 2$ and $p = 1$ has an explicit solution given by $\mathbf{s}_* = \text{sign}(\mathbf{s}_0)\ .*\ [|\mathbf{s}_0| - \alpha\mathbf{1}]_+$ or in scalar form $s_{j*} = \text{sign}(s_{j0})[|s_{j0}| - \alpha]_+$, where $\mathbf{s}_0 = \mathbf{A}^T\mathbf{x}$ is the minimum norm solution to $\mathbf{A}\,\mathbf{s} = \mathbf{x}$ without any constraints and $[x]_+ = \max\{0,\ x\}$.If specific sparsity profile is imposed $r = \text{card}(\mathbf{s}_*) < m$ a good heuristic is to solve the problem $EBPD$ for different values of $\alpha$, finding approximately the smallest value of $\alpha$ that satisfy desired constraint.

### 3.1. Fundamental Properties - Uniqueness and Optimal Solution for Sparse Representations

One fundamental question that is actually investigated by many researchers is to find sufficient conditions for a vector $\mathbf{x}$ to have a unique possibly sparsest and optimal representation as a linear combination of columns of the matrix $\mathbf{A}$ and to find such condition that the various heuristic and greedy algorithms such as BP, EBPD, FOCUSS or MP (Matching Pursuit) ensure optimal or close to optimal representations.[7, 41] Another important issue is establish equivalence conditions for various criteria and algorithms in the sense that obtained solutions have their non-zero components at the same locations and with the same signs.[7]

The this paper we present some fundamental properties and conditions which characterize optimal and unique solutions of sparse representation of signals. The following theorem provides a simple test to check whether the obtained results are optimal the sense that no significantly different linear expansion from the dictionary can provide both a smaller approximation error ($||\mathbf{e}||$) and a better sparsity.

Theorem 3.1. *Let us consider the system of linear underdetermined equations $\mathbf{A}\,\mathbf{s} = \mathbf{x}$ with $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $m < n$. Let $\mathbf{A}_r = \mathbf{A}_1 \in \mathbb{R}^{r\times m}$ designates the $r \leq m$ columns of the matrix $\mathbf{A}$ associated with $r$ non-zero elements of the desired vector $\mathbf{s}_*$. Furthermore, let a submatrix $\mathbf{A}_2 \in \mathbb{R}^{m\times(n-r)}$ designate the $(n-r)$ columns of the matrix $\mathbf{A}$ which are associated with the components of vector $\mathbf{s}_*$ equal to zero. If the pseudo-inverse matrix $\mathbf{A}_r^+ = \mathbf{A}_r^T(\mathbf{A}_r\mathbf{A}_r^T)^{-1}$ exists then the sparse representation $\mathbf{x} = \mathbf{A}_r\mathbf{s}_{1*} + \mathbf{e}_r$ is unique and optimal if and only if all the components of the vector*

$$\mathbf{g}_1 = \mathbf{A}_2^T\ [\mathbf{A}_r^T]^+\,\text{sign}(\mathbf{s}_{1*}) \tag{30}$$

*have magnitudes strictly less than one (i.e., $||\mathbf{g}_1||_\infty < 1$). In the case, when $\|\mathbf{g}_1\|_\infty > 1$ the solution is not optimal.*

It is worth to note that the vector $\mathbf{g}_1$ in (30) is rather insensitive to values of entries of $\mathbf{s}_{1*}$ since it depends only on their signs. Since we unlikely to know the optimal solution *a priori* the above theorem may look useless. However, using theorem 3.1, we can easily check whether the LP/BP or EBPD or any other heuristic algorithm, provides an optimal desired solution or not. In fact, the above Theorem has quite general nature and the condition (30) can be applied for many heuristic and greedy algorithms for subset selection such as BP, EBPD, MP, and FOCUSS algorithm.It should be also noted that the optimal solution $s_*$ of $(EBPD)$ problem can be equivalent to sparsest 0-norm solution $\mathbf{s}_0$ only in the sense that the both solutions has the non-zero components in the same positions and with same sign. This follows from the simple fact that $\mathbf{A}\,\mathbf{s}_* \neq \mathbf{x}$ for $(EBPD)$ due to nonzero regularization parameter $\alpha$. In order words, in order to obtain the sparsest possible solution $\alpha$ should be sufficiently small that $\mathbf{s}_*$ and $\mathbf{s}_0$ have their non-zero components at the same locations and with same signs, i.e., $\text{sign}(\mathbf{s}_0) = \text{sign}(\mathbf{s}_*)$.

THEOREM 3.2. *For any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and vector $\mathbf{x} \in \mathbb{R}^m$ with $m < n$, there exits a vector $\mathbf{s}_* \in \mathbb{R}^n$, which minimizes $J_1(\mathbf{s}) = \|\mathbf{s}\|_p$ subject to constraints $\mathbf{A}\,\mathbf{s} = \mathbf{x}$ with $p \leq 1$ such that the optimal vector $\mathbf{s}_*$ has at most $m$ non zero components. Furthermore, if the column vectors of the extended matrix $\bar{\mathbf{A}} = [\mathbf{x}\ \mathbf{A}] \in \mathbb{R}^{m\times(n+1)}$ satisfy the Haar condition then there exists a vector $\mathbf{s}_*$ which minimizes the $\|\mathbf{s}\|_1$ subject to the constraints $\mathbf{A}\,\mathbf{s} = \mathbf{x}$ that has exactly $m$ nonzero components.*

The above basic results can be extended and generalized for arbitrary $l_p$-norm or any norm non necessary differentiable one, if we apply the concepts of the subgradient and subdifferential. We say, that $\mathbf{g}$ is subgradient of function $J(\mathbf{x})$ if $J(\mathbf{y}) \geq J(\mathbf{x}) + \mathbf{g}^T(\mathbf{y} - \mathbf{x})\ \ \forall \mathbf{y}$. Set of all subgradients of $J(\mathbf{x})$ at $\mathbf{x}$ is called the subdifferential at $\mathbf{x}$ and it is denoted by $\partial J(\mathbf{x})$. For example, the subgradients of the $l_1$-norm can be formulated as the following sets $\partial\ \|\mathbf{e}\|_1 = \{\ \mathbf{g} \in \mathbb{R}^m\ :\ g_i = \text{sign}(e_i\ )\ \text{if}\ \ e_i \neq 0;\ \ |g_i| \leq 1\ \ \text{if}\ \ e_i = 0\}$.

We can formulate Theorem which characterizes solution of the $(EBPD)$ in more general form.

THEOREM 3.3. *The optimization problem $\min_{\mathbf{s}} \|\mathbf{x} - \mathbf{A}\ \mathbf{s}\|_{q_A} + \alpha\ \|\mathbf{s}\|_{p_S}$, where $\|\cdot\|_{q_A}$ and $\|\cdot\|_{p_S}$ are arbitrary norms has a sparse unique solution $\mathbf{s}_* \in \mathbb{R}^n$ with $\mathbf{e}(\mathbf{s}_*) = \mathbf{x} - \mathbf{A}\,\mathbf{s}_* \in \mathbb{R}^m$ if and only if exist $\mathbf{g}_{q_A} \in \partial\ \|\mathbf{x} - \mathbf{A}\,\mathbf{s}\|_{q_A}$ and $\mathbf{g}_{p_S} \in \partial\|\mathbf{s}\|_{p_S}$ such that*

$$-\mathbf{A}^T\mathbf{g}_{q_A} + \alpha\ \mathbf{g}_{p_S} = \mathbf{0}. \tag{31}$$

As special case, we can formulate the following important conditions characterizing the $l_1$-norm solution as: $\mathbf{s}_*$ is a unique sparse solution of the non-smooth optimization problem $\min_{\mathbf{s}} (\|\mathbf{x} - \mathbf{A}\ \mathbf{s}\|_1 + \alpha\ \|\mathbf{s}\|_1)$ if and only if there exists $\mathbf{g}_q \in \partial\|\mathbf{e}\|_1$ and $\mathbf{g}_p \in \partial\|\mathbf{s}\|_1$ which satisfies matrix equation: $-\mathbf{A}^T\mathbf{g}_q + \alpha\ \mathbf{g}_p = \mathbf{0}$, where components of $\mathbf{g}_p$ are determined as $g_{qi} = \text{sign}(e_i),\ \text{if}\ \ e_i = x_i - \mathbf{a}_i^T\mathbf{s}_* \neq 0$ otherwise $g_{qi}$ is arbitrary except $|g_{qj}| \leq 1$ and $g_{pj} = \text{sign}(s_{j*}),\ \ \text{if}\ \ s_{j*} \neq 0$, otherwise $g_{pj}$ is arbitrary except $|g_{pj}| \leq 1$.

## 4. ENHANCEMENT AND DENOISING OF MULTIDIMENSIONAL DATA

A conceptual model for the elimination of noise and other undesirable components from multi-sensory data is depicted in Figure 3. First, BSS or GCA are performed using suitably chosen robust (with respect to the noise) algorithm performing a signal decomposition described by a linear transformation of sensory data as $\mathbf{y}(k) = \mathbf{W}\mathbf{x}(k)$, where the vector $\mathbf{y}(k)$ represents the specific components (e.g., sparse, smooth, spatio-temporally decorrelated or statistically independent components). Then, the projection of interesting or useful or significant components $\tilde{y}_j(k)$ back onto the sensors level. The reconstructed or "cleaned" sensor signals are obtained by linear transformation $\hat{\mathbf{x}}(k) = \mathbf{W}^+\tilde{\mathbf{y}}(k)$, where $\mathbf{W}^+$ is some pseudo-inverse of the unmixing matrix $\mathbf{W}$ and $\tilde{\mathbf{y}}(k)$ is the vector obtained from the vector $\mathbf{y}(k)$ after removal of all the undesirable components (i.e., by replacing them with zeros). The entries of estimated attenuation matrix $\hat{\mathbf{A}} = \mathbf{W}^+$ indicate how strongly each sensor picks up each individual component. Back projection of some significant components $\hat{\mathbf{x}}(k) = \mathbf{W}^+\tilde{\mathbf{y}}(k)$ allows us not only remove some artifacts and noise but also to enhance recorded sensor data. In many cases the estimated components must be at first filtered or smoothed in order to identify all significant components.[27, 29, 42–44]

In addition to the denoising and artifacts removal, BSS techniques can be used to decompose sensor data into individual components, each representing a physically or physiologically distinct process or a source signal. The main idea here is to apply localization and imaging methods to each of these components in turn. The decomposition is usually based on the underlying assumption of sparsity and/or statistical independence between the activation of different sources involved.

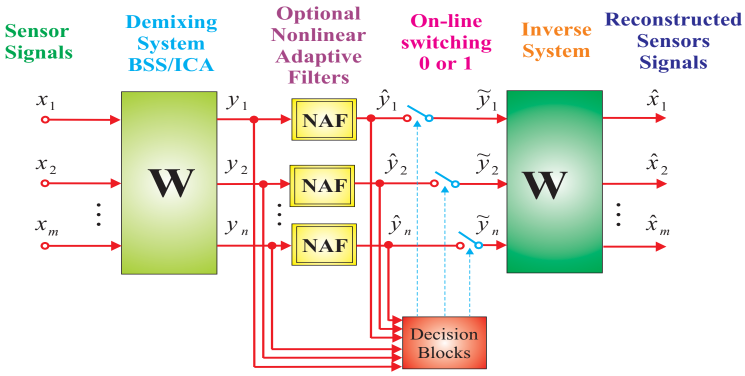

**Figure 3.** Basic models for removing undesirable components like noise and artifacts and enhancing multi-sensory (e.g., EEG/MEG) data using auxiliary nonlinear adaptive filters to smooth the extracted components and hard switches. Often the estimated components are also normalized, ranked, ordered and clustered in order to identify significant and physically meaningful sources or artifacts.

## 5. DISCUSSION AND CONCLUSIONS

In this paper we have discussed briefly several extensions and modifications of blind source separation and decomposition algorithms for spatio-temporal decorrelation, independent component analysis, and sparse component analysis where various criteria and constraints are imposed such linear predictability, smoothness, mutual independence, sparsity and non-negativity of extracted components. Especially, we described generalization and extension of ICA to MSD-ICA which relaxes considerably the condition on independence of original sources. Using these concepts in many cases, we are able to reconstruct (recover) the original sources and to estimate mixing and separating matrices, even if the original signals are not independent and in fact they are strongly correlated. Moreover, we propose a simple method for checking validity and true performance of BSS separation by applying the preprocessing with a bank of filters with various frequency characteristics.

## REFERENCES

1. S. Amari, A. Cichocki, and H. Yang, "A new learning algorithm for blind signal separation," in *Advances in Neural Information Processing Systems 1995*, M. C. M. David S. Touretzky and M. E. Hasselmo, eds., **8**, pp. 757–763, MIT Press: Cambridge, MA, 1996.
2. S. Amari and A. Cichocki, "Adaptive blind signal processing - neural network approaches," *Proceedings IEEE* **86**, pp. 1186–1187, 1998.
3. S. Amari and J.-F.Cardoso, "Blind source separation — semi-parametric statistical approach," *IEEE Trans. on Signal Processing* **45**, pp. 2692–2700, Dec. 1997.
4. A. Cichocki and S. Amari, *Adaptive Blind Signal And Image Processing (New revised and improved edition)*, John Wiley, New York, 2003.
5. A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley, New York, 2001.
6. D. D. Lee and H. S. Seung, "Learning of the parts of objects by non-negative matrix factorization," *Nature* **401**, pp. 788–791, 1999.
7. D. L. Donoho and M. Elad, "Representation via l1 minimization," *The Proc. National Academy of Science* **100**, pp. 2197–2202, March 2004.
8. H. H. Szu, P. Chanyagorn, and I. Kopriva, "Sparse coding blind source separation through powerline," *Neurocomputing* **48**, pp. 1015–1020, 2002.
9. H. H. Szu, S. Noel, S.-B. Yim, J. Willey, and J. Landa, "Multimedia authenticity protection with ica watermarking and digital bacteria vaccination," *Neural Networks* **16**, pp. 907–914, 2003.
10. M. Zibulevsky, P. Kisilev, Y. Zeevi, and B. Pearlmutter, "Blind source separation via multinode sparse representation," in *In Advances in Neural Information Processing Systems, (NIPS2001)*, pp. 185–191, Morgan Kaufmann, 2002.

11. F. Bach and M. Jordan, "Beyond independent components: trees and clusters," *Journal of Machine Learning Research* **4**, pp. 1205–1233, 2003.
12. L. Zhang, A. Cichocki, and S. Amari, "Multichannel blind deconvolution of nonminimum-phase systems using filter decomposition," *IEEE Transactions on Signal Processing* **52**(5), pp. 1430–1442, 2004.
13. S. Choi, A. Cichocki, and S. Amari, "Equivariant nonstationary source separation," *Neural Networks* **15**, pp. 121–130, 2002.
14. F. Miwakeichi, E. Martnez-Montes, P. A. Valds-Sosa, N. Nishiyama, H. Mizuhara, and Y. Yamaguchi, "Decomposing EEG data into space-time-frequency components using Parallel Factor Analysis," *NeuroImage* **22**(3), pp. 1035–1045, 2004.
15. A. Cichocki and R. Unbehauen, "Robust neural networks with on-line learning for blind identification and blind separation of sources," *IEEE Trans. Circuits and Systems I : Fundamentals Theory and Applications* **43**, pp. 894–906, Nov. 1996.
16. A. Cichocki, R. Unbehauen, and E. Rummert, "Robust learning algorithm for blind separation of signals," *Electronics Letters* **30**, pp. 1386–1387, August 1994.
17. A. Cichocki, R. Bogner, L. Moszczyński, and K. Pope, "Modified Hérault-Jutten algorithms for blind separation of sources," *Digital Signal Processing* **7**, pp. 80 – 93, April 1997.
18. S. Fiori, "A fully multiplicative orthoghonal-group ICA neural algorithm," *Electronics Letters* **39**(24), pp. 1737–1738, 2003.
19. S. A. Cruces, L. Castedo, and A. Cichocki, "Robust blind source separation algorithms using cumulants," *Neurocomputing* **49**, pp. 87–118, Dec. 2002.
20. S. Cruces, A. Cichocki, and L. Castedo, "An iterative inversion approach to blind source separation," *IEEE Trans. on Neural Networks* **11**(6), pp. 1423–1437, 2000.
21. S. A. Cruces and A. Cichocki, "Combining blind source extraction with joint approximate diagonalization: Thin algorithms for ICA," in *Proceedings of 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pp. 463–468, Riken, ICA, (Kyoto, Japan), Apr. 2003.
22. A. Bell and T. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computation* **7, no. 6**, pp. 1129–1159, Nov 1995.
23. A. Cichocki and R. Thawonmas, "On-line algorithm for blind signal extraction of arbitrarily distributed, but temporally correlated sources using second order statistics," *Neural Processing Letters* **12**, pp. 91–98, August 2000.
24. J. Stone, "Blind source separation using temporal predictability," *Neural Computation* **13**(7), pp. 1559–1574, 2001.
25. A. Cichocki, R. Thawonmas, and S. Amari, "Sequential blind signal extraction in order specified by stochastic properties," *Electronics Letters* **33**, pp. 64–65, Jan. 1997.
26. A. Cichocki, T. M. Rutkowski, and K. Siwek, "Blind signal extraction of signals with specified frequency band," in *Neural Networks for Signal Processing XII: Proceedings of the 2002 IEEE Signal Processing Society Workshop*, pp. 515–524, IEEE, (Martigny, Switzerland), Sept. 2002.
27. A. K. Barros and A. Cichocki, "Extraction of specific signals with temporal structure," *Neural Computation* **13**, pp. 1995–2000, September 2001.
28. H.-Y. Jung and S.-Y. Lee, "On the temporal decorrelation of feature parameters for noise-robust speech recognition," *IEEE Transactions on Speech and Audio Processing* **8**(7), pp. 407–416, 2000.
29. A. Cichocki and A. Belouchrani, "Sources separation of temporally correlated sources from noisy data using bank of band-pass filters," in *Third International Conference on Independent Component Analysis and Signal Separation (ICA-2001)*, pp. 173–178, (San Diego, USA), Dec. 9-13 2001.
30. R. R. Gharieb and A. Cichocki, "Second-order statistics based blind source separation using a bank of subband filters," *Digital Signal Processing* **13**, pp. 252–274, 2003.
31. A. Cichocki and P. Georgiev, "Blind source separation algorithms with matrix constraints," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* **E86-A**, pp. 522–531, Jan. 2003.

32. T. Tanaka and A. Cichocki, "Subband decomposition independent component analysis and new performance criteria," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP2004)*, **V**, pp. 541–544, (Montreal, Canada), May 2004.
33. A. Cichocki, Y. Li, P. G. Georgiev, and S. Amari, "Beyond ICA: Robust sparse signal representations," in *Proceedings of 2004 IEEE International Symposium on Circuits and Systems (ISCAS2004)*, **V**, pp. 684–687, (Vancouver, Canada), May 2004.
34. A. Cichocki, S. M. Amari, K. Siwek, T. Tanaka, and et al., "ICALAB toolboxes for signal and image processing *www.bsp.brain.riken.go.jp*," (JAPAN), 2004.
35. J. Himberg, A. Hyvärinen, and F. Esposito, "Validating the independent components of neuroimaging time series via clustering and visualization," *NeuroImage* **22**(3), pp. 1214–1222, 2004.
36. F. Mainecke, A. Ziehe, M. Kawanabe, and K.-R. Müller, "A resampling approach to estimate the stability of one dimensional or multidimensional independent components," *NeuroImage* **49**(13), pp. 1514–1525, 2002.
37. Y. Li, A. Cichocki, S. Amari, S. Shishkin, J. Cao, and F. Gu, "Sparse representation and its applications in blind source separation," in *Seventeenth Annual Conference on Neural Information Processing Systems (NIPS-2003)*, (Vancouver), Dec. 2003.
38. Y. Li, A. Cichocki, and S. Amari, "Analysis of sparse representation and blind source separation," *Neural Computation* **16**, pp. 1193–1204, June 2004.
39. P. G. Georgiev and A. Cichocki, "Sparse component analysis of overcomplete mixtures by improved basis pursuit method," in *Proceedings of 2004 IEEE International Symposium on Circuits and Systems (ISCAS2004)*, **V**, pp. 37–40, (Vancouver, Canada), May 2004.
40. F. J. Theis, P. G. Georgiev, and A. Cichocki, "Robust overcomplete matrix recovery for sparse sources using a generalized Hough transform," in *Proceedings of 12th European Symposium on Artificial Neural Networks (ESANN2004)*, pp. 223–232, (Bruges, Belgium), Apr. 2004.
41. K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Computation* **15**, pp. 349–396, February 2003.
42. A. Cichocki and S. Vorobyov, "Application of ICA for automatic noise and interference cancellation in multisensory biomedical signals," in *Proceedings of the Second International Workshop on ICA and BSS, ICA'2000*, pp. 621–626, (Helsinki, Finland), 19-22 June 2000.
43. S. Vorobyov and A. Cichocki, "Blind noise reduction for multisensory signals using ICA and subspace filtering, with applications to EEG analysis," *Biological Cybernetics* **86**, pp. 293–303, 2002.
44. A. Cichocki, R. Gharieb, and T. Hoya, "Efficient extraction of evoked potentials by combination of Wiener filtering and subspace methods," in *Proc. of IEEE Int. Conf. Acoustics, Speech, Signal Processing, ICASSP-2001*, pp. 3117–3120, (Utah, USA), May 7-11 2001.