Journal of Machine Learning Research 4 (2003) 1339-1364 Submitted 10/02; Published 12/03

A Multiscale Framework For Blind Separation of Linearly Mixed Signals

Pavel Kisilev PAULK@TX.TECHNION.AC.IL
Michael Zibulevsky MZIB@EE.TECHNION.AC.IL
Department of Electrical Engineering, Technion
Haifa 32000, Israel
Yehoshua Y. Zeevi ZEEVI@EE.TECHNION.AC.IL
Department of Electrical Engineering, Technion
Haifa 32000, Israel
and
Columbia University
New York, NY 10027, USA

Editors: Te-Won Lee, Jean-François Cardoso, Erkki Oja and Shun-ichi Amari

Abstract

We consider the problem of blind separation of unknown source signals or images from a given set of their linear mixtures. It was discovered recently that exploiting the sparsity of sources and their mixtures, once they are projected onto a proper space of sparse representation, improves the quality of separation. In this study we take advantage of the properties of multiscale transforms, such as wavelet packets, to decompose signals into sets of local features with various degrees of sparsity. We then study how the separation error is affected by the sparsity of decomposition coefficients, and by the misfit between the probabilistic model of these coefficients and their actual distribution. Our error estimator, based on the Taylor expansion of the quasi-ML function, is used in selection of the best subsets of coefficients and utilized, in turn, in further separation. The performance of the algorithm is evaluated by using noise-free and noisy data. Experiments with simulated signals, musical sounds and images, demonstrate significant improvement of separation quality over previously reported results.

Keywords: Blind Source Separation, Multiscale transforms, Maximum Likelihood, Wavelets

1. Introduction

In a variety of communication and signal sensing applications, crosstalk or mixing of several source signals occurs. The Blind Source Separation (BSS) is concerned with scenarios where an N-channel sensor signal $\mathbf{x}(\xi)$ is generated by M unknown source signals $s_m(\xi), m = 1,\ldots,M$, linearly mixed by an unknown $N \times M$ mixing, or crosstalk, matrix $\mathbf{A}$, and possibly corrupted by additive noise $\mathbf{n}(\xi)$:

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

The independent variable ξ represents either time, spatial coordinates in the case of images, spatio-temporal variables in the case of video sequences, or wavelength in the case of multispectral or other optical signals. BSS is concerned with estimation of the mixing matrix $\mathbf{A}$ and, thereby, solving the inverse problem of estimating the M-dimensional source signal $\mathbf{s}(\xi)$.

The assumption of statistical independence of the source components $s_m(\xi)$ lends itself to the Independent Component Analysis (ICA) (Bell and Sejnowski, 1995, Cardoso, 1998, Hyvärinen, 1999a, Pearlmutter and Parra, 1996, Hyvärinen, 1999b) and justified by the physics of many practical applications. A more powerful assumption is *sparsity* of the decomposition coefficients, when the signals are properly represented (Zibulevsky and Pearlmutter, 2001, Zibulevsky et al., 2001, 2002). Sparsity means that only a small fraction of coefficients differ significantly from zero. Let each $s_m(\xi)$ have a sparse representation of its decomposition coefficients c_{mk} obtained by means of the set of representation functions $\{\varphi_k(\xi)\}$:

$$s_m(\xi) = \sum_k c_{mk}\, \varphi_k(\xi). \tag{2}$$

The functions $\varphi_k(\xi)$ are called *atoms* or *elements* of the representation space that may constitute a basis or a frame. These elements do not necessarily have to be linearly independent and, instead, may form an overcomplete set (or dictionary), for example, wavelet-related dictionaries: wavelet packets, stationary wavelets, etc. (see, for example, Coifman and Wickerhauser, 1992, Mallat, 1998, Stanhill and Zeevi, 1996). The corresponding representation of the mixtures, according to the same signal dictionary, is:

$$x_m(\xi) = \sum_k^K y_{mk}\, \varphi_k(\xi), \tag{3}$$

where y_{mk} are the decomposition coefficients of the *mixtures*.

Often, overcomplete dictionaries, e.g. wavelet packets, contain bases as their subdictionaries, with the corresponding elements being orthogonal. Let us consider the case wherein the subset of functions $\{\varphi_k : k \in \Omega\}$ is constructed from the mutually orthogonal elements of the dictionary. This subset can be either complete or undercomplete, since only synthesis coefficients are used in the estimation of the mixing matrix, and sources are recovered without using these coefficients, as explained below. A more general case, wherein the above subset of functions is overcomplete, leads to the *maximum a posteriori approach* to the BSS problem (Zibulevsky et al., 2001). This approach arrives at a non-convex objective function which makes convergence not stable when optimization starts far from the solution (we use a more robust Maximum Likelihood formulation of the problem).

Let us define vectors $\mathbf{y}_k$ and $\mathbf{c}_k,\ k \in \Omega$ to be constructed from the k-th coefficients of mixtures and of sources, respectively. From (1) and (2), and using the orthogonality of the subset of functions $\{\varphi_k : k \in \Omega\}$, the relation between the decomposition coefficients of the mixtures and of the sources, when the noise is small, is

$$\mathbf{y}_k \approx \mathbf{A}\mathbf{c}_k.$$

Note, that the relation between decomposition coefficients of the mixtures and the sources is exactly the same relation as in the original domain of signals, where $\mathbf{x}_\xi \approx \mathbf{A}\mathbf{s}_\xi$. Then, estimation of the mixing matrix and of sources is performed using the decomposition coefficients of the mixtures $\mathbf{y}_k$ instead of the mixtures $\mathbf{x}_\xi$.

The property of sparsity often yields much better source separation than standard ICA, and can work well even with more sources than mixtures (Bofill and Zibulevsky, 2001). In many cases there are distinct groups of coefficients, wherein sources have different sparsity properties. The proposed multiscale, or multiresolution, approach to the BSS is based on selecting only a subset of features, or coefficients, $\mathbf{Y} = \{\mathbf{y}_k : k \in \Omega\}$, which is best suited for separation, with respect to the sparsity of coefficients and to the separability of sources' features. In our experiments we use the Wavelet Packet (WP) transform that reveals the structure of signals, wherein several subsets of the WP coefficients have significantly better sparsity and separability than others.

We also investigate how the separation error is affected by the sparsity of a particular subset of the decomposition coefficients, and by the misfit between the probabilistic model of these coefficients and their actual distribution. Since the actual separation errors are not tractable in practice, we propose to use a method for estimation of the separation error, based on the Taylor expansion of the quasi log-likelihood function. The obtained estimates are used for selection of the best subset of coefficients, i.e. the subset that leads to the lowest estimated error. After the new data set is formed from this best subset, one uses it in the separation process, that can be accomplished by any of the standard ICA algorithms or by clustering.

The performance of our approach is verified on noise-free and noisy data. Our experiments with 1D signals and images demonstrate that the proposed method significantly improves separation quality, as compared to the results obtained by using sparsity of a complete set of decomposition coefficients.

2. Sparse Source Separation

Sparse sources can be separated by each one of several techniques, for example, by approaches based on the maximum likelihood (ML) considerations (Pham et al., 1992, Cardoso, 1997, Pearlmutter and Parra, 1996), or by approaches based on geometric considerations (Puntonet et al., 1995, Prieto et al., 1998, Pajunen et al., 1996). In the former case, the algorithm estimates the *unmixing* matrix $\mathbf{W} = \mathbf{A}^{-1}$, while in the later case the output is the estimated mixing matrix. In both cases, these matrices can be estimated only up to a column permutation and a scaling factor (Comon et al., 1991).

2.1 Quasi Maximum Likelihood BSS

In this section, we discuss the ML and the quasi ML solution of the BSS problem, based on the data in the domain of decomposition coefficients. The term *quasi* indicates that the proposed *hypothetical* density is used instead of the true one (see discussion below).

Let $\mathbf{Y}$ be the *features*, or new data matrix of dimension $M \times K$, where K is the number of features, or data points, and the coefficients $\mathbf{y}_k$'s form the columns of $\mathbf{Y}$. Note, that, in general, the rows of $\mathbf{Y}$ can be formed from either the samples of sensor signals, that is, mixtures, or, as in our setting, from their decomposition coefficients. In the latter case, $\mathbf{Y} = \left\{\mathbf{y}_k : k \in \mathbf{\Omega}_{jn}\right\}$, where $\mathbf{\Omega}_{jn}$ are the subsets indexed on the WP tree, as explained in the section on the Multinode analysis. We are interested in the maximum likelihood estimate of $\mathbf{A}$ given the data $\mathbf{Y}$.

2.1.1 QUASI LOG-LIKELIHOOD FUNCTION

We assume that the coefficients c_{mk} are i.i.d. random variables with the joint probability density function (pdf)

$$f_{\mathbf{C}} = p(\mathbf{C}) = \prod_{m,k} p(c_{mk}),$$

where $p(c_{mk})$ is of an exponential type:

$$f_{c_{mk}} = p(c_{mk}) = N_q \exp\{-\nu(c_{mk}, q)\}.$$

The normalization constant N_q is omitted in the further calculations, since it has no effect on the maximization of the log-likelihood function. In a particular case wherein

$$\nu(c_{mk},q) = |c_{mk}|^q/q,$$

and $q<1$, the above distribution is widely used for modeling sparsity (Lewicki and Sejnowski, 2000, Olshausen and Field, 1997). For $q=0.5\div 1$, it approximates rather well the empirical distributions of wavelet coefficients of natural signals and images (Buccigrossi and Simoncelli, 1999). A smaller q corresponds to a distribution with greater sparsity.

Let $\mathbf{W}\equiv\mathbf{A}^{-1}$ be the unmixing matrix to be estimated. Taking into account that $\mathbf{Y}=\mathbf{AC}$, we arrive at the standard expression of the ICA log-likelihood, but with respect to the decomposition coefficients:

$$L_{\mathbf{W}}(\mathbf{Y}) = K\log|\det\mathbf{W}| - \sum_{m=1}^{M}\sum_{k=1}^{K}\nu([\mathbf{WY}]_{mk},q).$$

In the case wherein $\nu(\cdot,q)=|\cdot|^q/q$, the second term in the above log-likelihood function is not convex for $q<1$, and non-differentiable, and, therefore, is difficult to optimize. Furthermore, the parameter q of the true pdf is usually unknown in practice, and estimation of this parameter along with the estimation of the unmixing matrix is a difficult optimization problem. Therefore, it is convenient to replace the actual $\nu(\cdot)$ with its *hypothetical* substitute, a smooth, convex approximation of the absolute value function, for example $\tilde{\nu}(c_{mk})$=$\sqrt{c_{mk}^2+\zeta}$, with ζ being a smoothing parameter. This approximation has a minor effect on the separation performance, as indicated by our numerical results. The corresponding *quasi* log-likelihood function is

$$\tilde{L}_{\mathbf{W}}(\mathbf{Y}) = K\log|\det\mathbf{W}| - \sum_{m=1}^{M}\sum_{k=1}^{K}\tilde{\nu}([\mathbf{WY}]_{mk}). \tag{4}$$

2.1.2 NATURAL GRADIENT ALGORITHM UPDATE

Maximization of $\tilde{L}_{\mathbf{W}}(\mathbf{Y})$ with respect to $\mathbf{W}$ can be solved efficiently by several methods, for example the Natural Gradient (NG) algorithm (Cichocki et al., 1994, Amari et al., 1996) or, equivalently, by the Relative Gradient (Cardoso and Laheld, 1996), as implemented in the ICA/EEG Matlab toolbox (Makeig, 1998).

The derivative of the quasi log-likelihood function with respect to the matrix parameter $\mathbf{W}$:

$$\frac{\partial\tilde{L}_{\mathbf{W}}(\mathbf{y})}{\partial\mathbf{W}} = (K\mathbf{I}-\tilde{\psi}(\mathbf{c})\mathbf{c}^T)(\mathbf{W}^T)^{-1}, \tag{5}$$

where $\mathbf{c}$ is the 'column stack' version of $\mathbf{C}$, and $\tilde{\psi}(\mathbf{c}) = [\tilde{\psi}_1(c_1)...\tilde{\psi}_{MK}(c_{MK})]^T$, where $\tilde{\psi}_{mk}(c_{mk})\equiv -(\log\tilde{f}_{c_{mk}})'$ are the so-called *score functions.* Note, that, in our case, $\tilde{\psi}_{mk}(c_{mk})=\tilde{\nu}'(c_{mk})$. The learning rule of the Natural Gradient algorithm is given by

$$\Delta\mathbf{W} = \frac{\partial\tilde{L}_{\mathbf{W}}(\mathbf{y})}{\partial\mathbf{W}}\mathbf{W}^T\mathbf{W} = (K\mathbf{I}-\tilde{\psi}(\mathbf{c})\mathbf{c}^T)\mathbf{W}.$$

In the original Natural Gradient algorithm, as implemented in (Makeig, 1998), the above update equation is expressed in terms of the *non-linearity*, which has the form of the cumulative density function (cdf) of the hypothetical source distribution. The built-in non-linearity is the logistic function, $\tilde{g}_s = 1/(1+\exp(-s))$. The relation of the non-linearity to the score function, is as follows: since $\tilde{f}_c = \tilde{g}'_c$, we have, by differentiation:

$$\tilde{\psi}(c) \equiv -(\log \tilde{f}_c)' = -\frac{\tilde{f}'_c}{\tilde{f}_c} = -\frac{\tilde{g}''_c}{\tilde{g}'_c}.$$

The score function for the above $\tilde{g}_s$, is $\tilde{\psi}_g = -1 + 2\tilde{g}_s$. The corresponding function $\tilde{\nu}(\cdot)$ is, as before, a kind of smooth approximation of the absolute value function, with a smoothing parameter of order 1. In order to adapt the algorithm to our purposes, we use hypothetical density with $\tilde{\nu}(c_{mk})$ $= \sqrt{c_{mk}^2 + \zeta}$, and the corresponding score function is $\tilde{\psi}(c_{mk}) = c_{mk}/\sqrt{c_{mk}^2 + \zeta}$.

2.2 Clustering-based BSS

In the case of geometry-based methods, separation of sparse sources can be achieved by clustering along orientations of data concentration in the N-dimensional space wherein each column $\mathbf{y}_k$ of the matrix $\mathbf{Y}$ represents a data point and N is the number of mixtures. Let us consider a two-dimensional noiseless case, wherein two source signals, $s_1(t)$ and $s_2(t)$, are mixed by a 2×2 matrix $\mathbf{A}$, arriving at two mixtures $x_1(t)$ and $x_2(t)$. In this case, the data matrix is constructed from these mixtures $x_1(t)$ and $x_2(t)$). An example of a scatter plot of 2 mixtures $x_1(t)$ versus $x_2(t)$, constructed from 2 sparse sources, is shown in Figure 1.

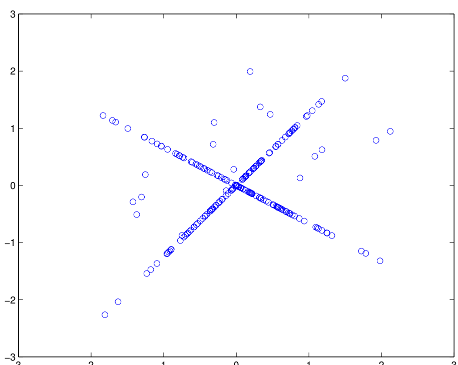


Figure 1: An example of a scatter plot of 2 mixtures $x_1(t)$ versus $x_2(t)$ of 2 sparse sources.

If only one source, say $s_1(t)$, was present, the sensor signals would be

$$\begin{array}{rcl} x_1(t) & = & a_{11}s_1(t) \\ x_2(t) & = & a_{21}s_1(t) \end{array}.$$

All the data points of the scatter diagram x_2 versus x_1 cluster in this case co-linearly along the straight line defined by the vector $[a_{11}a_{21}]^T$. A similar highly structured distribution results when two *sparse* natural sources are present simultaneously. An example of fragments of 2 sparse sources is shown in Figure 2.

In this sparse case, at each particular instant where a sample of the first source is large, there is a high probability, that the corresponding sample of the second source is small, and the sampling point

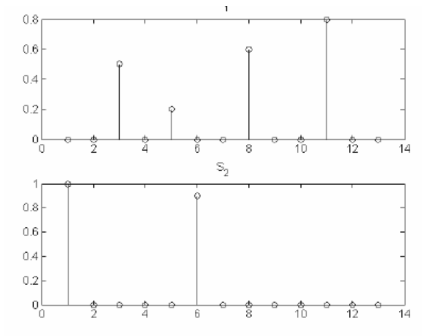


Figure 2: An example of fragments of 2 sparse sources.

projected onto the scatter diagram lies close to the corresponding straight line. The same arguments apply to the complementary points, dominated by the second source. As a result, data points are densely clustered along two dominant orientations, which are directly related to the columns of **A**.

Source signals are rarely sparse in their native original domain. An example of fragments of 2 *not* sparse sources is shown in Figure 3. An example of a scatter plot of 2 mixtures $x_1(t)$ versus $x_2(t)$, constructed from 2 *not* sparse sources, is shown in Figure 4.

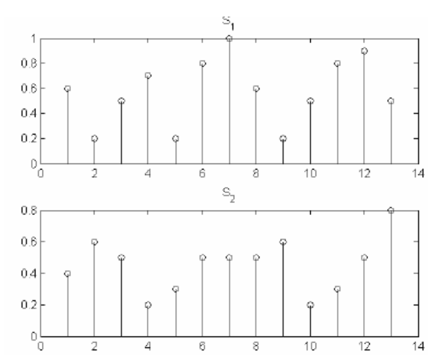


Figure 3: An example of fragments of 2 *not* sparse sources.

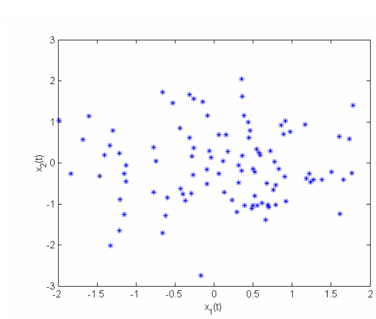


Figure 4: An example of a scatter plot of 2 mixtures $x_1(t)$ versus $x_2(t)$ of 2 *not* sparse sources.

In contrast to the source samples in their original domain, the decomposition coefficients of sources, Equation (2), are much more sparser. We therefore construct the data matrix $\mathbf{Y}$ from the decomposition coefficients of the mixtures, Equation (3), rather than from the original mixtures.

In order to determine orientations of scattered data, we project the data points onto the surface of a unit sphere by normalizing corresponding vectors, and then apply a standard clustering algorithm. This clustering approach is applicable even in cases where the number of sources exceeds the number of sensors. Under such circumstances, separated clusters consist of samples of scaled versions of sources. An example of scatter plot of 2 mixtures of 6 sparse sources is shown in Figure 5. Note, that in case of natural signals that are not sparse in their native domain (say time, wavelength, position or functions of any other independent variable), a proper projection onto a space of sparse representation will yield a highly structured scatter plot of the decomposition coefficients of the mixtures, similar to the one shown in Figure 5.

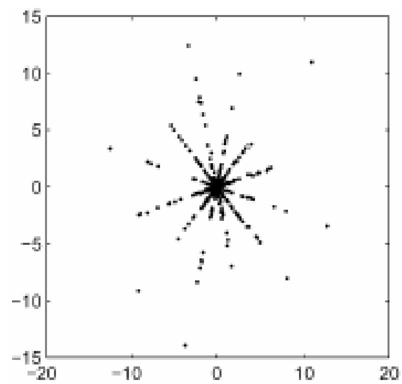


Figure 5: An example of scatter plot of 2 mixtures of 6 sparse sources.

The proposed *clustering procedure* used in our BSS algorithm is as follows:

1. Form the feature matrix $\mathbf{Y}$, by inserting samples of the sensor signals or (*subset of*) their decomposition coefficients into the corresponding rows of the matrix;

2. Normalize feature vectors: $\mathbf{y}_k = \mathbf{y}_k / \|\mathbf{y}_k\|_2$, in order to project data points onto the surface of a unit sphere, where $\|\cdot\|_2$ denotes the l_2 norm;

Before normalization, it is reasonable to remove data points with a very small norm, since these are very likely crosstalk-corrupted by small coefficients from other sources.

3. Move data points to a half-sphere, *e.g.* by forcing the sign of the first coordinate y_k^1 to be positive: IF $y_k^1 < 0$ THEN $\mathbf{y}_k = -\mathbf{y}_k$;

Without this operation, each set of clustered-along-a-line data points would yield two clusters on opposite sides of the sphere.

4. Estimate cluster centers by using a clustering algorithm. The coordinates of these centers will form the columns of the estimated mixing matrix $\tilde{\mathbf{A}}$;

We used *Fuzzy C-means* (FCM) clustering algorithm (Bezdek, 1981) as implemented in the Matlab Fuzzy Logic Toolbox.

2.3 Sources Recovery

The estimated unmixing matrix $\hat{\mathbf{W}} = \hat{\mathbf{A}}^{-1}$ can be obtained by either clustering, or by the quasi-ML *approach, implemented along with the Natural Gradient* (we call it simply *Natural Gradient*), or by other algorithms, which are applied to either the complete data set, or to some subsets of data (see the subsequent section). In any case, this matrix and, therefore, the sources, can be estimated only up to a column permutation and a scaling factor (Comon et al., 1991). The sources are recovered in their original domain by

$$\hat{\mathbf{s}}(t) = \hat{\mathbf{W}}\mathbf{x}(t).$$

It should be stressed, that if the above clustering approach is used, the estimation of mixing matrix and of sources is *not* restricted to the case of *square* mixing matrices, although the sources recovery is more complicated in the rectangular case.

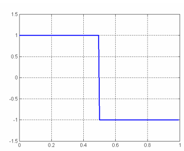
3. Multiscale BSS

To provide intuitive insight into the practical implications of our main ideas, we first discuss an example of separation of 1D *block signals*, that are piecewise constant, with random amplitude and duration of each constant piece (Figure 6).

3.1 Motivating Example: Sparsity of Random Blocks in the Haar Basis Space

It is well known, that the Haar wavelet basis provides compact representation of block functions. Let take a close look at the Haar wavelet coefficients at different resolution levels *j=0,1,...,J*. Wavelet basis functions at the finest resolution level *j=J* are obtained by translation of the Haar mother wavelet:

$$\varphi_{j=J}(t) = \begin{cases} 1 & \text{if } 0 \leqslant t < \frac{1}{2} \\ -1 & \text{if } \frac{1}{2} \leqslant t < 1 \\ 0 & \text{otherwise} \end{cases}$$



Taking the scalar product of a function $s(t)$ with the wavelet $\varphi_J(t-\tau)$, results in a finite differentiation of the function $s(t)$ at the point $t=\tau$. This implies that the number of non-zero coefficients at the finest resolution for a block function will roughly correspond to the number of jumps of this function. Proceeding to the next, coarser resolution level, we have the wavelet $\varphi_{J-1}(t) = \{\frac{1}{2}, \text{if } 0 \leqslant t < 1; -\frac{1}{2}, \text{ if } 1 \leqslant t < 2; 0 \text{ otherwise}\}$. The number of non-zero coefficients at this level still corresponds to the number of jumps, but the total number of coefficients at this level is halved, and so is the sparsity. Proceeding to coarser resolutions, we encounter levels where the support of a wavelet $\varphi_j(t)$ is comparable to the typical distance between jumps in the function $s(t)$. In this case, most of the coefficients are expected to be nonzero, and, therefore, sparsity fades away.

To demonstrate how this influences the accuracy of BSS, we randomly generate two block-signal sources (Figure 6, two upper plots.), and mix them by the crosstalk matrix

$$\mathbf{A} = \begin{pmatrix} 0.8321 & 0.6247 \\ -0.5547 & 0.7809 \end{pmatrix}.$$

Resulting sensor signals, or mixtures, $x_1(t)$ and $x_2(t)$ are shown in the two lower plots of Figure 6. The scatter plot of $x_1(t)$ versus $x_2(t)$ does not exhibit any visible distinct orientation of clustering

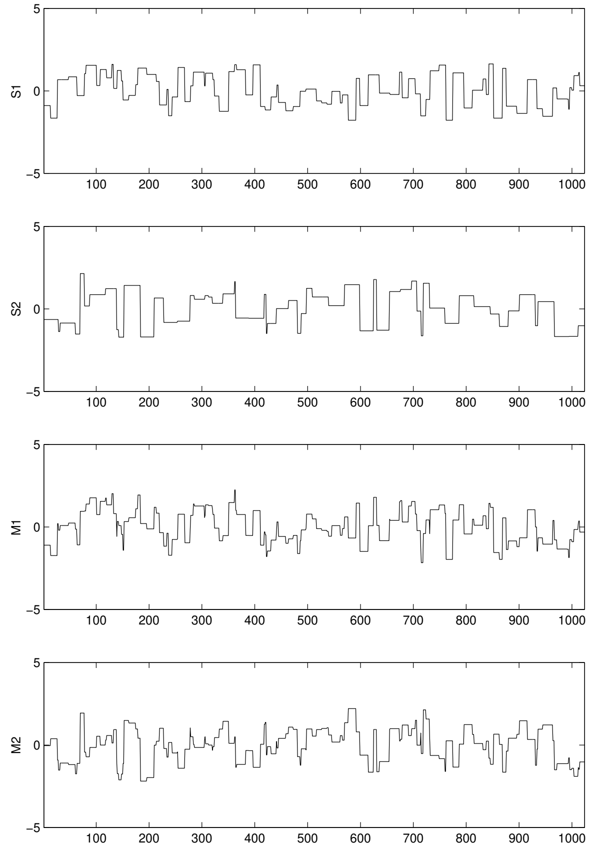


Figure 6: Random block signals (two upper) and their mixtures (two lower).

(Figure 7, left). Similarly, in the scatter plot of all the wavelet coefficients, the two clusters' orientations are hardly detectable (Figure 7, middle). In contrast, the scatter plot of the wavelet coefficients at the highest level of resolution (Figure 7, right) depicts two distinct orientations, which correspond to the columns of the mixing matrix.

Since a cross-talk matrix $\mathbf{A}$ is estimated only up to a column permutation and a scaling factor, in order to measure the separation accuracy, we normalize the original sources $s_m(t)$ and their *corresponding* estimated sources $\tilde{s}_m(t)$. The averaged (over sources) normalized squared error (NSE) is then computed as:

$$NSE = \frac{1}{M}\sum_{m=1}^{M}(\|\hat{s}_m - s_m\|_2/\|s_m\|_2)^2. \tag{6}$$

In the noise-free case, this error is equal to the averaged residual *cross-talk error* (CTE):

$$CTE = \frac{1}{M}\sum_{m=1}^{M}\frac{\sum_{l=1,}^{M}(\hat{\mathbf{A}}^{-1}\mathbf{A})_{ml}^2 - (\max\{(\hat{\mathbf{A}}^{-1}\mathbf{A})_m\})^2}{\sum_{l=1,}^{M}(\hat{\mathbf{A}}^{-1}\mathbf{A})_{ml}^2},$$

where $\max\{(\hat{\mathbf{A}}^{-1}\mathbf{A})_m\}$ is the largest element in the m-th row of the matrix $\hat{\mathbf{A}}^{-1}\mathbf{A}$.

Resulting separation errors for block sources are presented in the lower part of Figure 7. The largest error (11%) is obtained on the raw data, and the smallest (0.002%) – on the wavelet coeffi-

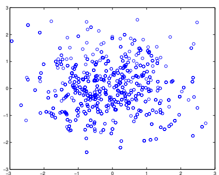
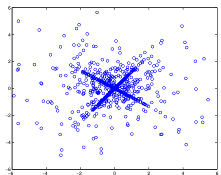
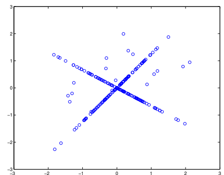
	Raw signals	All wavelet coefficients	High resolution WT coefficients
			
NG	10.9	0.056	0.0024
FCM	10.2	0.04	0.002

Figure 7: Separation of mixed block signals: scatter plots of sensor signals (left), and of their wavelet coefficients (middle and right). Lower rows present the normalized mean-squared separation error (%) corresponding to the Natural Gradient (NG), and to the Fuzzy C-Means (FCM) clustering, respectively.

cients at the highest resolution, which have the best sparsity. Using all wavelet coefficients yields intermediate sparsity and performance.

3.2 Multinode Representation

Our choice of a particular wavelet basis and of the sparsest subset of coefficients was obvious in the above example: it was based on knowledge of the structure of piecewise constant signals. For sources having oscillatory components (like sounds or images with textures), other systems of bases functions, such as wavelet packets (Mallat, 1998, Chen et al., 1998), trigonometric function libraries (Wickenhauser, 1994), or multiwavelets (Weitzer et al., 1997) might be more appropriate. In particular, the wavelet packet library consists of the triple-indexed family of functions:

$$\varphi_{j,i,n}(t) = 2^{j/2}\varphi_n(2^j t - i),\ j,i \in \mathbf{Z},\, n \in \mathbf{N}, \tag{7}$$

where j,i are the scale and shift parameters, respectively, and n is the frequency-like parameter. [Roughly speaking, n is proportional to the number of oscillations of a mother wavelet $\varphi_n(t)$]. These functions form a binary tree whose nodes are indexed by the depth of the level, j, and the node number $n = 0,1,2,3,...,2^j - 1$ at the specified level j. The same indexing is applied to corresponding subsets of wavelet packet coefficients (Figure 8), and is used for the scatter diagrams in the section on experimental results.

3.3 Adaptive Selection of Data Subsets for Separation

As noted, the choice of a particular wavelet basis and of the sparsest subset of coefficients is pretty obvious in the context of the example with block signals. The representations of signals mentioned in the section on the multinode analysis is usually considered in the context of a one-to-one mapping. In particular, in the case of the wavelet packets representation, the best basis is chosen from the library of bases, and is used for a decomposition, providing a complete set of features (coefficients). Another strategy, the basis pursuit (Chen et al., 1998), chooses the wavelet representation resulting in the smallest l_1-norm of the coefficients.

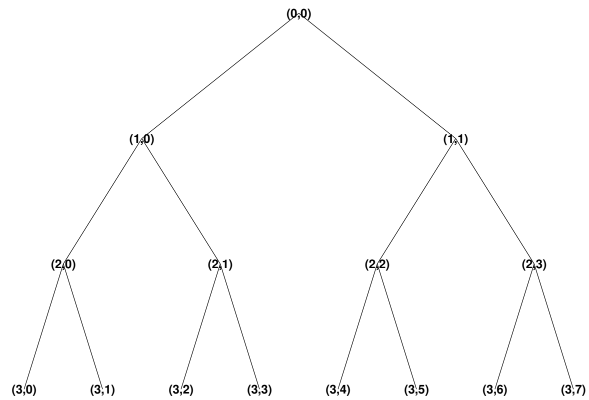


Figure 8: Wavelet Packets tree of subsets of coefficients, indexed by scale (first index) and frequency-like (second index) parameters pair, according to (7).

In contrast, in the context of the signal source separation problem, it is useful to extract first an overcomplete set of features. An example of overcomplete set of WP features of *mixtures* of two Flute signals is presented in Figure 9. In this figure, the tree of scatter plots is formed by the subsets of WP coefficients which 'live' in the corresponding nodes of the WP tree. The WP analysis reveals an almost noise-free subset of coefficients–the most left scatter plot in the bottom row. This set is best suited for further separation.

Generally, in order to construct an overcomplete set of features, our approach to the BSS allows to combine multiple representations (for example, wavelet packets with various generating functions and various families of wavelets, along with DFT, DCT, etc.). After this overcomplete set is constructed, we choose the 'best' subsets (with respect to some separation criteria) and form a new set used for separation. This set can be either complete or undercomplete.

3.3.1 HEURISTIC SELECTION OF BEST NODES

it is usually difficult to decide in advance which nodes, i.e. subsets of data, contain the sparsest sets of coefficients. Furthermore, there may be a situation wherein only one, or, more generally, not all of the sources are represented in one of the subspaces (at a particular node). In this case, the corresponding scatter plot will reveal only one or several dominant orientation. Such a node can still be very useful for separation, provided the points on the scatter plot are well clustered.

One can apply the following heuristic approach for choosing the appropriate nodes (Zibulevsky et al., 2002). First, for every node of the tree, we apply our clustering algorithm, and compute a measure of clusters' distortion. In our experiments we used a standard *global distortion*, the mean squared distance of data points to the centers of their own, closest, clusters (here again, the weights of the data points can be incorporated):

$$d = \sum_{k=1}^{K} \min_{m} \| u_m - y_k \|_2,$$

where K is the number of data points, u_m is the m-th centroid coordinates, y_k is the k-th data point coordinates, and $\|.\|_2$ is the l_2 norm, which is a sum-of-squares distance. Second, we choose a few best nodes with the minimal distortion, combine their coefficients into one data set and apply

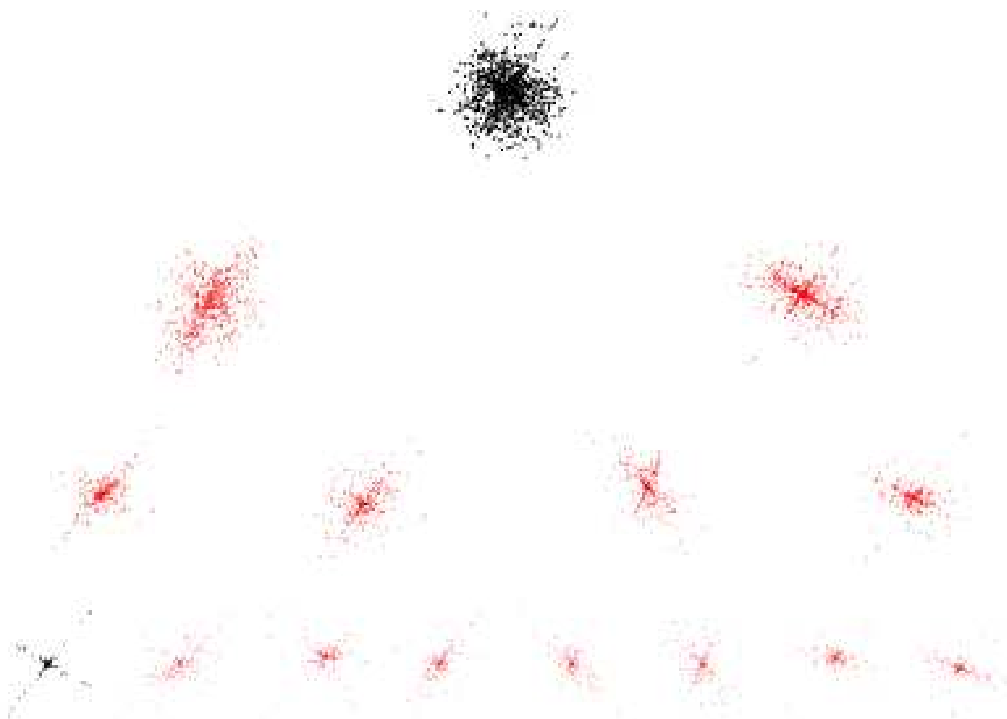


Figure 9: Wavelet Packets (WP) based multinode data structure: the example with mixture of 2 Flutes. Each node depicts scatter plot of WP coefficients of the sensor signals, wherein the coefficients are taken from the corresponding nodes of the WP tree (Figure 8).

a separation algorithm, clustering or Natural Gradient, to these data. In the case of the wavelet packets, where two sets of children basis functions span the same subspace as their parent, we should check that there is no redundancy in the set of the best nodes, so that no children and their parents are present at the same time in the final set used for separation. For images, there are four children subsets of functions for each parent set, and therefore, each set of parent coefficients is split into four subsets of coefficients: approximations, horizontal, vertical and diagonal details.

Alternatively, the following iterative approach can be used. First, we apply the Wavelet Transform to original data, and apply a standard separation technique to these data in the transform domain. As such separation technique, we can use either the Natural Gradient, our clustering approach, or simply apply some optimization algorithm to minimize the corresponding log-likelihood function. This provides us with an initial estimate of the unmixing matrix $\mathbf{W}$ and with the estimated source signals. Then, at each iteration of the algorithm, we apply a multinode representation (for example, WP, or trigonometric library of functions) to the estimated sources, and calculate a measure of sparsity for each subset of coefficients. For example, one can consider the $l_1$norm of the coefficients $\sum_{m,k} |c_{mk}|$, its modification $\sum_{m,k} \log |c_{mk}|$, or some other entropy-related measures.

Finally, we combine a new data set from the subsets with the highest sparsity (in particular, we can take for example, the best 10% of coefficients), and apply some separation algorithm to the new data. The iteration of the algorithm is completed. This process can be repeated till convergence is achieved.

3.3.2 ERROR ESTIMATOR

When signals have a complex nature, the heuristic approach may not be as robust as desired, and the error-related statistical quantities must be estimated. We use the following approach. First, we apply the Wavelet Transform to original data, and apply a standard separation technique (Natural Gradient, clustering, or optimization algorithm to minimize the log-likelihood function) to these data in the transform domain. Second, given the initial estimate of $\mathbf{W}$, and the subsets of data (coefficients of mixtures) for each node, we estimate the corresponding error variance, as described below. Finally, we choose a few best nodes (or, simply, the best one) with small estimated errors, combine their coefficients into one data set, and apply a separation algorithm to these data. In the rest of this section, we focus on the issues related to the estimation of error variance.

An estimate of the error covariance matrix can be derived using the second order Taylor expansion of the log-likelihood function. Let $\mathbf{W}_* = \mathbf{A}^{-1}$ be the exact solution of the BSS problem (1); it satisfies

$$\mathbf{W}_* = \arg\max E\left[\tilde{L}_{\mathbf{W}}(\mathbf{Y})\right].$$

Further, let $\mathbf{W}_o$ be the estimate of $\mathbf{W}_*$, based on the particular realization of data, $\mathbf{Y}_o$, that is,

$$\mathbf{W}_o = \arg\max \tilde{L}_{\mathbf{W}}(\mathbf{Y}_o).$$

Note, that $\nabla\tilde{L}_{\mathbf{W}_o}(\mathbf{Y}_o) = 0$, while $\nabla\tilde{L}_{\mathbf{W}_*}(\mathbf{Y}_o) \neq 0$ (though, $E\left[\nabla\tilde{L}_{\mathbf{W}_*}(\mathbf{Y})\right] = 0$). We want to estimate the error, $\Delta\mathbf{W} = \mathbf{W}_* - \mathbf{W}_o$.

Using the linearization of $\nabla\tilde{L}_{\mathbf{W}}(\mathbf{Y}_o)$ around $\mathbf{W}_o$

$$\nabla\tilde{L}_{\mathbf{W}_o+\Delta\mathbf{W}}(\mathbf{Y}_o) \simeq \nabla\tilde{L}_{\mathbf{W}_o}(\mathbf{Y}_o) + \nabla^2\tilde{L}_{\mathbf{W}_o}(\mathbf{Y}_o)\Delta\mathbf{W}$$

we obtain

$$\Delta\mathbf{W} \simeq \mathbf{H}^{-1}\nabla\tilde{L}_{\mathbf{W}_*}(\mathbf{Y}_o),$$

where $\mathbf{H} = \nabla^2\tilde{L}_{\mathbf{W}_o}(\mathbf{Y}_o)$ is the Hessian. Therefore, the error covariance matrix, whose diagonal elements correspond to the error variances of sources, can be approximated as

$$E\left[\Delta\mathbf{W}\Delta\mathbf{W}^T\right] \simeq \mathbf{H}^{-1}\Sigma\left(\mathbf{H}^{-1}\right)^T, \tag{8}$$

where

$$\Sigma = E\left[\nabla\tilde{L}_{\mathbf{W}_*}(\mathbf{Y})\nabla\tilde{L}_{\mathbf{W}_*}(\mathbf{Y})^T\right]$$

is the covariance matrix of the gradient vector.

Further, denote by $\nabla\tilde{L}_k$ the column stack vector of the k-th data point gradient

$$\nabla\tilde{L}_k = \frac{\partial\tilde{L}_{\mathbf{W}}(\mathbf{c}_k)}{\partial\mathbf{W}} = (\mathbf{I} - \tilde{\psi}(\mathbf{c}_k)\mathbf{c}_k^T)(\mathbf{W}^T)^{-1},$$

obtained from (5). Then, since

$$\nabla\tilde{L}_{\mathbf{W}}(\mathbf{Y}) = \sum_{k=1}^{K}\nabla\tilde{L}_k,$$

the covariance matrix Σ can be estimated from K data points as:

$$\hat{\Sigma} = K\hat{\Gamma},$$

where

$$\hat{\Gamma} = \frac{1}{K}\sum_{k=1}^{K}\left[\nabla\tilde{L}_k \nabla\tilde{L}_k^T\right].$$

The Hessian matrix can be calculated either analytically, or via numerical evaluation (see Kisilev et al., 2002,for details). In the latter case, its j-th column is approximated by:

$$H_j(\mathbf{w}) \simeq \frac{\nabla L(\mathbf{w}+\xi\mathbf{e}_j) - \nabla L(\mathbf{w})}{\xi},$$

where $\mathbf{w}$ is a column stack version of $\mathbf{W}$, ξ is a small constant and $\mathbf{e}_j = [0,0,...,1,0,...,0]^T$ is the vector with the only nonzero entry at the j-th position.

The diagonal element $\hat{\varepsilon}_{ii}^2$ in the error covariance matrix above (8) represents an estimate of the squared error introduced to the reconstructed i-th source by cross-talks from other sources. The analytic expression for the mean square relative contamination of the reconstructed i-th source by the j-th source, for the quasi-ML estimator, is given by Pham and Garat (1997):

$$\varepsilon^2 = \frac{1}{K}\frac{\alpha_i^2\alpha_j^2(\beta_j^2/\alpha_j^2 + \beta_i^2 - 2\beta_i^2\beta_j^2/\alpha_j)}{(1-\alpha_i\alpha_j)^2\beta_i^2\beta_j^2}, \tag{9}$$

where K is the number of data samples, and, parameters α and β are dependant on the source-related expectations. In our case, α and β are *coefficient-related:*

$$\alpha_i = \frac{E[\tilde{\psi}_i(\mathbf{c}_i)\mathbf{c}_i]}{E[\tilde{\psi}_i'(\mathbf{c}_i)]E[\mathbf{c}_i^2]};\ \beta_i = \frac{E[\tilde{\psi}_i(\mathbf{c}_i)\mathbf{c}_i]}{\sqrt{E[\tilde{\psi}_i^2(\mathbf{c}_i)]E[\mathbf{c}_i^2]}}. \tag{10}$$

Note, that the expectation operator $E[\cdot]$ is with respect to the *true* pdf $f_{\mathbf{C}}$, while $\tilde{\psi}(\mathbf{c}) = -(\log\tilde{f}_{\mathbf{C}})'$ is dependent on the hypothetical pdf $\tilde{f}_{\mathbf{C}}$.

4. Experiments

For reasons given in the previous section, the hypothetical density is used in the expression of the log-likelihood function (4). In our experiments we use the function $\nu(c_{mk},\zeta) = \sqrt{c_{mk}^2+\zeta}$, as a smooth approximation of the absolute value function.

4.1 Numerical Results: Theoretic and Estimated Separation Errors

In the following experiment, we study the effect of using the quasi log-likelihood instead of the likelihood function on the theoretical error performance. For this purpose, we evaluate expectations in (10) via numerical calculation of the corresponding integrals. Then, for various values of the smoothness parameter ζ, we calculate the error variance according to (9) with the sample size $K = 10^4$, as a function of the true density shape parameter, q^*. The results of this calculation are presented in Figure 10. Note, that the error variance drops dramatically as the sparseness of sources increases. This figure shows also that using hypothetical density has a minor effect on the theoretical error.

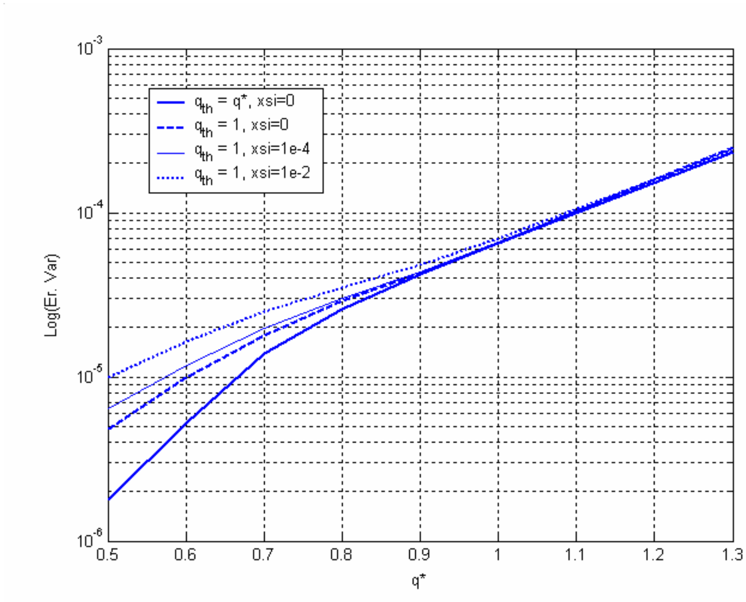


Figure 10: Theoretical cross-talk error variance: Using the true density, i.e. $q_{th} = q^*$ (solid bold line), vs. hypothetical densities with parameters $q_{th} = 1$, $\zeta = 0$, 10^{-2}, 10^{-4} for the ML estimator.

4.1.1 SYNTHETIC EXPONENTIAL SOURCES

In the following experiment, we generate two source signals with 10^4 samples each, drawn from the exponential distribution $f_s \sim \exp\{-|s|^q/q\}$ with $q^* = 0.5$. These two signals are mixed together with a 2x2 matrix of normally distributed random numbers. Consequently, the matrix is normalized for the purpose of the error calculation. In order to separate these signals, we apply the following algorithms: 1) the original Natural Gradient with the built-in non-linearity, as implemented in the ICA/EEG Matlab toolbox (Makeig, 1998); 2) the modified Natural Gradient with the non-linearity corresponding to our quasi log-likelihood function, with parameters $q_{th} = 1$, and $\zeta = 10^{-4}$, that is, using $\nu(c_{mk}, \zeta) = \sqrt{c_{mk}^2 + 10^{-4}}$; 3) the Matlab function *fminu* (an implementation of the BFGS Quasi-Newton optimization method (Bertsekas, 1999)), applied to optimize our quasi log-likelihood function with the above parameter values. The following error-related quantities were calculated: 1) the error variance $\hat{\varepsilon}^2$, estimated from data according to (9) by evaluation of corresponding expectations via averaging; 2) the error variances calculated according to (8); these are evaluated for each one of the above three algorithms; 3) actual squared separation errors of the algorithms; 4) theoretical error variance, calculated as described in the beginning of the Section 4.1. Tables 1 and 2 summarizes the results of the above experiment; the corresponding values are averaged over the trials. The smallest actual separation error is achieved by using the *fminu* function. This result is quite expected, since, generally speaking, a batch mode optimization algorithm (such as the BFGS Quasi-Newton) outperforms an online mode optimization algorithm (such as the Natural Gradient). The modified Natural Gradient outperforms the original one, as expected, since its hypothetical pdf is 'closer' to the true pdf of sources. Also, the error estimate (8) is closer to the theoretical error variance than the estimate according to (9).

	fminu	Mod. NG	Orig. NG
Actual sq. error	8e-5	1.2e-4	4.1e-4

Table 1: Actual squared cross-talk errors

	fminu
Actual sq. error	8e-5
Error var. estimate (9)	1.6e-4
Error var. estimate (8)	7.8e-5
Theoretical error variance	6.6e-5

Table 2: Estimates of error

4.1.2 NATURAL SOURCES

In the following experiment we apply the *fminu* function (as described above) to estimate **W** from each one of 22 data subsets formed by the WP coefficients of mixture of natural images. In Figure 11, we depict the following quantities, for each one of the 22 data subsets: the actual squared separation error and its estimate according to (8). The first subset corresponds to the complete set of the Wavelet Transform (WT) coefficients; the other subsets are indexed on the WP tree. Our 'error predictor' provides quite accurate estimates of the actual errors. Note, that for the best, the 5-th, subset, the separation error and its estimate are smaller by several orders, as compared to the corresponding quantities of the 1st set, the complete set of the WT coefficients.

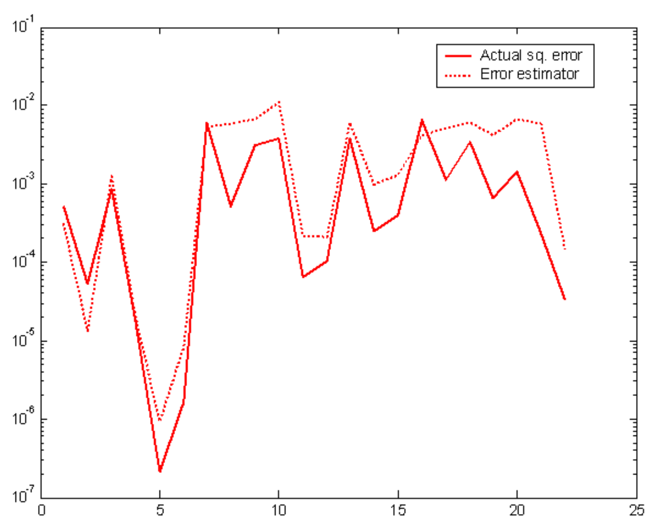


Figure 11: The actual squared cross-talk error [-], and its estimate (8) [..], evaluated for each one of 22 data sets formed by the corresponding subsets of WP coefficients.

4.2 Separation of Simulated and Real Signals and Images

The proposed blind separation method based on the wavelet-packet representation, was evaluated by using several types of signals. We have already discussed the relatively simple example of a random block signal. The second type of signal is a frequency modulated (FM) sinusoidal signal. The carrier frequency is modulated by either a sinusoidal function (FM signal) or by random blocks (BFM signal). The third type is a musical recording of flute sounds. Finally, we apply our algorithm to images. An example of such images is presented in Figure 12. Source images and their mixtures are shown at the upper two sets of plots, and the estimated images are shown in the lower two plots.

Figure 12: Applying Multiscale BSS method to noise-free mixtures: two source images (upper pair), their mixtures (middle pair) and separated images (lower pair).

In order to compare accuracy of our multiscale BSS method with that attainable by standard methods, we form the following feature sets: (1) raw data, (2) Short Time Fourier Transform (STFT) coefficients (in the case of 1D signals), (3) Wavelet Transform coefficients (4) Wavelet packet coefficients at the best nodes found by the proposed error estimator (8), while using various wavelet families with different smoothness: Daubechies family - db-1 (or, Haar), db-4, and db-8 mother wavelets. The higher number indicates higher smoothness. In the case of image separation, we used the Discrete Cosine Transform (DCT) instead of the STFT, and the Symmlet (almost symmetric) wavelet family - sym-4 and sym-8 mother wavelets instead of db-4 and db-8, when using wavelet transform and wavelet packets.

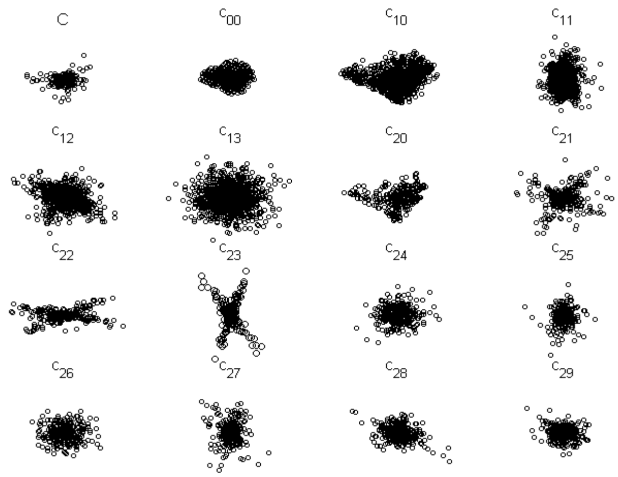


Figure 13: Scatter plots of the wavelet packet (WP) coefficients of mixtures of two images; subsets are indexed on the WP tree.

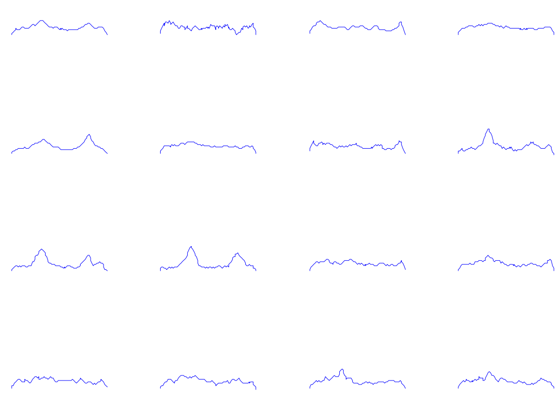


Figure 14: Distributions of angles (orientations) characterizing the scatter diagrams of the WP coefficients of mixtures of two images.

4.2.1 UNDERSTANDING SCATTER PLOT AND PDF DIAGRAMS

Let us consider an example of image separation from two mixtures of two sources (Figure 12). Figure 13 shows corresponding scatter plots of the wavelet packet coefficients of mixtures. These scatter plots correspond to the various nodes of the wavelet packet tree. The upper left scatter plot, marked with 'C', corresponds to the complete set of coefficients at all nodes. The rest are the scatter plots of sets of coefficients indexed on a wavelet packet tree. Generally speaking, the more distinct the two dominant orientations appear on these plots, the more precise is the estimation of the mixing matrix, and, therefore, the better is the quality of separation. Note, that only two nodes (the most left ones in the second from the bottom row) show clear orientations. These nodes will most likely be selected by the algorithm for further estimation process.

Figure 14 shows distributions of angles (orientations) formed by points on the corresponding scatter plots of the wavelet packet coefficients at various nodes. Here, the more distinct peaks assure better separation.

4.2.2 SEPARATION FROM NOISE-FREE MIXTURES

Table 3 summarizes results of experiments in which we applied the (original) Natural Gradient to each noise-free feature set. In the case of the WP features, the best subset was selected, using the proposed error estimator (8). In these experiments, we compared the quality of separation of deterministic signals by calculating NSE's, or residual crosstalk errors, according to (6). In the case of random block and BFM signals, we performed 100 Monte-Carlo simulations and calculated the normalized mean-squared errors (NMSE) for the above feature sets.

Signals	raw data	STFT	WT db8	WT haar	WP db8	WP haar
Blocks	10.16	2.669	0.174	0.037	0.073	0.002
BFM sine	24.51	0.667	0.665	2.34	0.2	0.442
FM sine	25.57	0.32	1.032	6.105	0.176	0.284
Flutes	1.48	0.287	0.355	0.852	0.154	0.648
Images	raw data	DCT	WT sym8	WT haar	WP sym8	WP haar
	4.88	3.651	1.164	1.114	0.36	0.687

Table 3: Normalized squared cross-talk errors [%]: Applying Natural Gradient-based separation to raw data and decomposition coefficients in various domains. In the case of wavelet packets (WP), the best nodes, selected by our error predictor, were used.

From Table 3 it is clear that using our best nodes method, implemented with the proposed error estimator, outperforms all other feature sets, including complete set of wavelet coefficients, for each type of signals. Similar improvements were achieved by using the FCM algorithm along with the heuristic data subset selection. In the case of the random block signals, using the Haar wavelet function for the wavelet packet representation yields a better separation than using some smooth wavelet, e. g. 'db-8'. The reason is that these block signals, that are not natural signals, have a sparser representation in the case of the Haar wavelets. In contrast, as expected, natural signals such as the Flute's signals are better represented by smooth wavelets, that in turn provide a better separation. This is another advantage of using sets of features at multiple nodes along with various families of 'mother' functions: one can choose best nodes from several decomposition trees simultaneously.

4.2.3 SEPARATION FROM NOISY MIXTURES

In order to verify the performance of our method in presence of noise, we added various types of noise (white Gaussian and salt&pepper) to three mixtures of three images at various SNR's. Table 4 summarizes these experiments in which we applied our approach along with the separation via the modified Natural Gradient algorithm.

Signal-to-noise energy ratio		100	30	10	3
Mixtures w. white gaussian noise	CTE	0.31	0.56	2.05	13.72
	NSE	0.39	1.69	12.25	54.29
Mixtures with salt&pepper noise	CTE	0.32	1.09	4.15	21.69
	NSE	0.39	3.11	16.81	69.88

Table 4: Performance of the algorithm in presence of various sources of noise in mixtures: Normalized mean-squared (NSE) and cross-talk (CTE) errors for image separation, applying our multiscale adaptive approach along with the Natural Gradient based separation.

An example of 'difficult' source separation from noisy mixtures with Gaussian noise is shown in Figure 15. It turns out that the ideas used in wavelet based signal denoising (see for example work by Donoho (1995) and references therein), are applied to signal separation from *noisy mixtures*. In particular, in case of white Gaussian noise, the noise energy is uniformly distributed over all wavelet coefficients at various scales. Therefore, at sufficiently high signal-to-noise energy ratios (SNR), the large coefficients of the signals are only slightly distorted by the noise coefficients. As a result, the presence of noise has a minor effect on the estimation of the unmixing matrix (see the *CTE* entries in Table 4). Note, that the *NSE* entries reflect the noise energy passed to the reconstructed sources from the mixtures. Our algorithm provides reasonable separation quality (CTE of about 4%) for SNR's of about 10 and higher. On the contrast, the Natural Gradient algorithm applied to the noisy mixtures themselves, failed completely to separate source images, arriving at CTE of 47% even in the case of SNR=30. We should stress here that, although our adaptive best nodes method performs reasonably well even in the presence of noise, it is not supposed to further denoise the reconstructed images. The post-filtering can be achieved by some denoising method, after initial source signals are separated. For example, in Figure 15, a simple wavelet denoising method from (Donoho, 1995) was applied to separated images.

5. Discussion and Conclusions

Although the problem of BSS has been dealt with extensively in the context of linearly mixed signals with no additional effects of the medium such as convolution, one should further extend and improve the state of the art in this research since there are enough important cases where the linear mixing model provides a good approximation of the physics of the problem. This is the

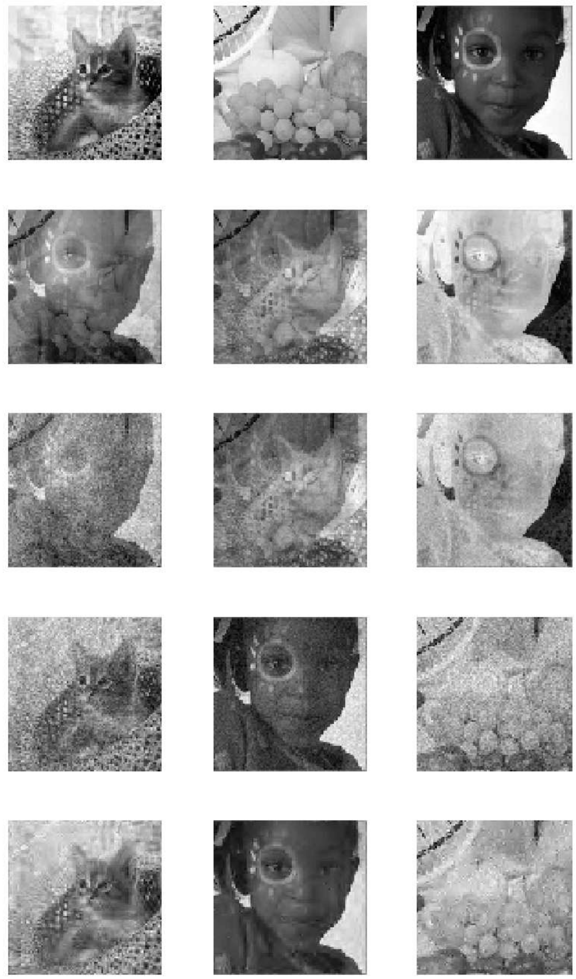


Figure 15: Applying Multiscale BSS to noisy mixtures: three source images (1st row), their mixtures (2nd row), the mixtures with additive white Gaussian noise (3d row), separated images (4th row), and post-filtered separated images (5th row)

case, for example, in the superposition of reflections on an image observed through a semireflective medium (Farid and Adelson, 1999, Bronstein et al., 2003). Although the optical system can not be considered in this case to be spatially invariant, and the simple linear mixing model provides only a reasonably good approximation of the problem, the localized multiscale geometrical approach yields excellent results that are by far better than those obtained by the conventional ICA algorithms. One reason for the improvement in this case is that the multinode representation permits 'locking on' a sparse data subset that can be considered to represent best the result of a locally spatial invariant mixing operation. Examples of (approximately) linear mixing-only in the time domain occur in cross talk between twisted pairs of telephone lines and similarly coupled transmission media. The linear mixing model, without any convolutive effects, provides also a powerful paradigm for the interpretation of a variety of medical signals and images such as optical imaging of normal and malignant tissue, where the signals measured from each group can be considered to reflect mixtures of optical signatures of molecules that are specific to the two groups but at different ratios, as well as of signatures of molecules that are common to both types of tissue. The latter case, to which we devote a great deal of our current effort, represents a problem of unmixing more sources than measured types of signals. Indeed, this type of a problem can not be solved by the ICA or ICA-based algorithms. It can be solved, however, by our approach of clustering of co-linearly scattered data in the space of the coefficients, obtained by projecting the mixtures into the space of sparse representation.

The proposed method improves the separation quality by utilizing the structure of signals projected onto a proper space, wherein certain subsets of the wavelet packet coefficients depict significantly better sparsity and, therefore, contribute to better separability than others. Other multiresolution representations based on wavelet-type (Coifman and Wickerhauser, 1992), nonseparable two-dimensional wavelets (Weitzer et al., 1997), and multiwindow Gabor-type frames (Zibulski and Zeevi, 1997) can most likely provide better results in specific cases of image subspaces. The approach depicted in this study using the specific example of wavelet packets can then be adapted to the other representations in a straightforward manner, by constructing the appropriate multinode data structure. In addition, the block partitioning of the image structure into subsets of projected data, can also be incorporated into the proposed multinode framework of BSS, to better exploit the localized spatial-invariant properties of images as well as those of the imaging system; properties that are not valid in the case of most signals and images because they are in most cases nonstationary. Nevertheless, the local stationarity is a reasonable assumption in most cases. Of the variety of wavele-type and similar representations that combine the properties of localization, scaling and additional parameters that are specific to each of the representations, there exist an optimal representation for each specific signal or image subspace. Selection or designing of the optimal representation is not, however, a straightforward problem. Although each of the wavelet-type transforms provides a sparse representation of most of the natural signals or images, there major differences in sparsity and clustering in the scatter diagrams can be observed when one uses different representations. In the example of the block signals, it is obvious that Haar wavelets are as close to optimal as one can probably get. Indeed, these wavelets provided much better estimation of the mixing matrix and thereby approximately one fifth of the cross talk error obtained by using smooth (db8) wavelets. The difference between the separation results obtained by using these types of wavelets is further substantiated when the wavelets are casted within the framework of wavelet packets. The squared cross-talk error obtained by using the Haar WP is thirty five times smaller than the error obtained by using the db8-WP, and almost two orders of magnitude smaller than the error obtained by using the conventional (Table 2). If the results obtained in the case of using the Haar-WP for estimation of the mixing matrix and separation of the mixed signals may be considered to be a benchmark, then indeed the match between the natural types of signals and images, used as examples in this study, is far from being optimal. The smallest squared cross-talk error obtained in the case of the separating mixed two flutes' signals by using the db8-WP is approximately two orders of magnitude larger than the 'benchmark' result. It is, however, one fourth of error obtained by using the Haar-WP. Thus, whereas the Haar-WP is by far much better than the db8-WP in the case of separating block signals, it is the other way around in the case of flute signals. Indeed the physics of the flute sounds constrains its time varying signals to be smooth, and it is therefore better to use smooth wavelets. Since the structure of images is less understood, it is even more difficult to come up with a recipe for constructing the optimal space of sparse representation. Based on some experience with nonseparable multiwavelets (Weitzer et al., 1997), we expect that such representations will provide better (closer to optimal) results. Yet, one has to identify the right mother wavelets that are most suitable for the representation of specific subspaces of images.

Simulation results and experiments with mixtures of both one- and two-dimensional natural signals substantiate our assertion that sparsity of the decomposition coefficients has a major effect on quality and efficiency of the BSS. Using the hypothetical log-likelihood function, that is, a smooth approximation of the likelihood derived from the actual pdf of sources, has an insignificant effect on the performance. The estimator of the error variance, based on the Taylor expansion of the quasi log-likelihood function, facilitates efficient selection of the best subset(s) of coefficients and, in fact, provides a reasonable metric for the vaguely defined concept of optimal selection of subsets of coefficients. Once this subset is selected, the mixing matrix is estimated using only this new subset of coefficients, by either clustering approach or by maximizing log-likelihood with some optimization algorithm (e.g., by the Natural Gradient).

It should be stressed here that the proposed approach to BSS can be applied in the context of any higher dimensional problem, such as volumetric medical or other natural data, wherein the number of sources is larger than the number of mixtures. In this case, sparse distribution of the properly transformed (projected) mixtures will cluster co-linearly in the higher dimensional scattered data space, and each colinear cluster will serve as an estimator of the elements of one of the source signal. This geometric (clustering) approach simplifies significantly the process of BSS and permits, as noted, the solution of the ill-posed inverse problems in cases that can not be dealt with by the ICA-based approach.

Acknowledgments

We thank Arkadi Nemirovski for useful discussions. This work has been supported in part by the Ollendorff Minerva Center, by the HASSIP research network program HPRN-CT-2002-00285, sponsored by the European Comission, and by the Israeli Ministry of Science. It is currently supported also by the Medical Imaging Group, Dept. of Biomedical Eng. Columbia University and by the ONR-MURI Grant N000M-01-1-0625.

References

S. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems 8*. MIT Press, 1996.

Anthony J. Bell and Terrence J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

Dimitri P. Bertsekas. *Nonlinear Programming: Second Edition*. Athena Scientific, Belmont, Massachussets, 1999.

J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

Pau Bofill and Michael Zibulevsky. Underdetermined blind source separation using sparse representation. *Signal Processing*, 81(11):2353–2362, 2001.

A. M. Bronstein, M. M. Bronstein, M. Zibulevsky, and Y. Y. Zeevi. Blind separation of reflections using sparse ica. *Proc. 4th International Symposium on Independent Component Analysis*, pages 227–232, 2003.

R. W. Buccigrossi and E. P Simoncelli. Image compression via joint statistical characterization in the wavelet domain. *IEEE Transactions on Image Processing*, 8(12):1688–1701, 1999.

Jean-François Cardoso. Infomax and maximum likelihood for blind separation. *IEEE Signal Processing Letters*, (4):112–114, 1997.

Jean-François Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 9 (10):2009–2025, October 1998.

Jean-François Cardoso and Beate Laheld. Equivariant adaptive source separation. *IEEE Transactions on Signal Processing*, 44(12):3017–3030, 1996.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998.

A. Cichocki, R. Unbehauen, and E. Rummert. Robust learning algorithm for blind separation of signals. *Electronics Letters*, 30(17):1386–1387, 1994.

R. R. Coifman and M. V. Wickerhauser. Entropy-based algorithms for best-basis selection. *IEEE Transactions on Information Theory*, 38:713–718, 1992.

P. Comon, C. Jutten, and J. Herault. Blind separation of sources, part ii: Problem statement. *Signal Processing*, 24:11–20, 1991.

D. L. Donoho. De-noising by soft thresholding. *IEEE Transactions on Information Theory*, 41(3): 613–627, 1995.

H. Farid and E. H. Adelson. Separating reflections from images using independent component analysis. *JOSA*, 17(9):2136–2145, 1999.

A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999a.

A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, (2):94–128, 1999b.

International Conference on Neural Information Processing, Hong Kong, September 24–27 1996. Springer-Verlag.

P. Kisilev, M. Zibulevsky, and Y. Y. Zeevi. Subset selection in multiscale blind source separation. Technical report, CCIT Report No 512, Technion. Haifa, Israel, 2002.

M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2):337–365, 2000.

Scott Makeig. ICA toolbox for psychophysiological research. Computational Neurobiology Laboratory, the Salk Institute for Biological Studies, 1998. http://www.cnl.salk.edu/˜ ica.html.

S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1998.

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37:3311–3325, 1997.

P. Pajunen, A. Hyvrinen, and J. Karhunen. Non-linear blind source separation by self-organizing maps. In ICONIP'96 ICO (1996), pages 1207–1210.

Barak A. Pearlmutter and Lucas C. Parra. A context-sensitive generalization of ICA. In ICONIP'96 ICO (1996), pages 151–157.

D.T. Pham and P Garat. Blind separation of a mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Transactions on Signal Processing*, 45(7):1712–1725, 1997.

D.T. Pham, P. Garrat, and C. Jutten. Separation of a mixture of independent sources through a maximum likelihood approach. In *European Signal Processing Conference*, pages 771–774, 1992.

A. Prieto, C. G. Puntonet, and B. Prieto. A neural algorithm for blind separation of sources based on geometric prperties. *Signal Processing*, 64(3):315–331, 1998.

C. G. Puntonet, A. Prieto, C. Jutten, M. Rodriguez-Alvarez, and J. Ortega. Separation of sources: A geometry-based procedure for reconstruction of n-valued signals. *Signal Processing*, 46(3): 267–284, 1995.

D. Stanhill and Y. Y. Zeevi. 2d multiwavelets for image representation. *The 19th Convention of the IEEE*, pages 251–254, 1996.

D. Weitzer, D. Stanhill, and Y. Y. Zeevi. Nonseparable two-dimensional multiwavelet transform for image coding and compression. *Proc. SPIE*, 3309:944–954, 1997.

M. V. Wickenhauser. *Adapted wavelet analysis: from theory to software*. Wellesley, MA, A K Peters, Ltd., 1994.

M. Zibulevsky, P. Kisilev, Y. Y. Zeevi, and B. A. Pearlmutter. Blind source separation via multinode sparse representation. In *Advances in Neural Information Processing Systems 12*. MIT Press, 2002.

M. Zibulevsky, B. A. Pearlmutter, P. Bofill, and P. Kisilev. Blind source separation by sparse decomposition. In S. J. Roberts and R. M. Everson, editors, *Independent Components Analysis: Princeiples and Practice*. Cambridge University Press, 2001.

Michael Zibulevsky and Barak A. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural Computations*, 13(4):863–882, 2001.

Meir Zibulski and Yehoshua Y. Zeevi. Analysis of multi-window gabor-type schemes by frame methods. *Applied and Computational Harmonic Analysis*, 4:188–221, 1997.