BLIND SEPARATION OF MORE SOURCES THAN MIXTURES USING SPARSITY OF THEIR SHORT-TIME FOURIER TRANSFORM

Pau Bofill *

Dept. d'Arquitectura de Computadors
Universitat Politècnica de Catalunya
Visiting CARTAH, Univ. of Washington
pau@ac.upc.es

Michael Zibulevsky †

Department of Computer Science
University of New Mexico
Albuquerque, NM 87131
michael@cs.unm.edu

ABSTRACT

This paper focuses on *underdetermined* Blind Source Separation, that is, the separation of N sources from M linear mixtures when $M < N$. We exploit the *sparsity* of the short-time Fourier transform when applied to music and speech signals. Given the mixing matrix, a sparse representation of the sources is obtained by solving a low-dimensional linear programming (LP) problem for each of the data points *independently*. For $M = 2$ we propose a *shortest path*, closed-form solution to this LP problem that represents each data point as a linear combination of the pair of directions that *enclose* it. The mixing matrix can be estimated either by maximizing a likelihood function, using the above LP optimization at the internal step, or with a *clustering* algorithm that gives a much faster solution. In this work, for $M = 2$ we use a clustering algorithm based on a triangular potential function which infers both the mixing matrix and the number of sources.

Several experiments involving music and speech signals are described, including the separation of six sources from two mixtures.

1. BLIND SOURCE SEPARATION WITH MORE SOURCES THAN MIXTURES

Let $\mathbf{x}^t$ be an M-dimensional column vector corresponding to the output of M sensors at a given discrete time instant t, and let $\mathbf{X}$ be an $M \times T$ matrix corresponding to the sensor data at all times $t = 1, \ldots, T$ (i.e., row i of $\mathbf{X}$, denoted $\mathbf{X}_{i\cdot}$, corresponds to the i-th mixture signal). Let $\mathbf{S}$ be the $N \times T$ matrix of underlying source signals and let $\mathbf{A}$ be the $M \times N$ mixing matrix. The problem of *blind source separation* [6], in the noiseless case, consists of finding the solution to the following system of equations

$$\mathbf{X} = \mathbf{AS} \tag{1}$$

when $\mathbf{A}$ and $\mathbf{S}$ are *unknown*. For our purposes, a useful formulation of this system is obtained by decomposing $\mathbf{A}$ into its columns $\mathbf{a}^j$ and expanding for every data point

$$\mathbf{x}^t = \sum_{j=1}^{N} \mathbf{a}^j s_j^t, \quad \text{for } t = 1, \ldots, T. \tag{2}$$

Then, in the M-dimensional mixture space, the $\mathbf{a}^j$'s are vectors indicating the *directions* of the sources and the s_j^t's are the components of the data points in those directions. Since results are not affected by reciprocal rescaling of the $\mathbf{a}^j$'s and the s_j^t's, without loss of generality the $\mathbf{a}^j$ will be hereafter assumed to be normalized to unit length.

For $M = N$, several approaches to *Independent Component Analysis* have been used in the literature (see for instance [5] for a recent survey) to numerically solve equation (1) while assuming only statistical independence of the source components s_j^t. Of particular interest to the work presented here is the so-called *sparse* case, in which only a small number of the s_j^t's differ significantly from zero. Sparsity is often modeled by a Laplacian distribution [10].

One of the most popular ICA approaches is the InfoMax algorithm [2]. When specialized for the Laplacian distribution, it leads to the following objective function

$$\min_{\mathbf{W}} -T \log|\det \mathbf{W}| + \sum_{jt} |\mathbf{WX}|_{jt}, \tag{3}$$

with $\mathbf{W}$ being the estimate of $\mathbf{A}^{-1}$ and $(\mathbf{WX})_{jt} = s_j^t$ the estimates of the source components.

*With support from ajuts BE98/99, DGR-Generalitat de Catalunya.

†Supported by NSF CAREER award 97-02-311, the National Foundation for Functional Brain Imaging, an equipment grant from Intel corporation, the Albuquerque High Performance Computing Center, a gift from George Cowan, and a gift from the NEC Research Institute.

The drawback of the above formulation is that it assumes the existence of the inverse matrix $\mathbf{W}$. Therefore, it is unsuitable for the *underdetermined* case $M < N$. The alternative is to formulate the search in mixing space rather than separation space. Generalizing equation (1) to the case with additive Gaussian noise $\mathbf{X} = \mathbf{AS} + \mathbf{V}$, a maximum *a posteriori* log-probability analysis leads to the following objective function [10, 12]

$$\min_{\mathbf{A},\mathbf{S}} \frac{1}{2\sigma^2}||\mathbf{AS} - \mathbf{X}||^2 + \sum_{jt} |s_j^t|, \tag{4}$$

with σ^2 the variance of the noise $\mathbf{V}$. The first term corresponds to the sum squared reconstruction error, and the second term is the penalty for non-sparsity.

2. ESTIMATING THE MIXING MATRIX AND THE SOURCES SEPARATELY

As opposed to the case of a square mixing matrix, where finding $\mathbf{W}$ amounts to solving the problem $\mathbf{S} = \mathbf{WX}$, in the underdetermined case we are faced with *two* interrelated problems: estimating the mixing matrix $\mathbf{A}$ *and* estimating the sources $\mathbf{S}$. Trying to solve both of them at the same time as in equation (4) is a difficult multivariate optimization problem.

Yet if we assume that the matrix $\mathbf{A}$ is given, the problem of finding the sources can be formulated *independently* for each data point x^t, leading to T tractable small problems

$$\min_{\mathbf{s}^t} \frac{1}{2\sigma^2}||\mathbf{A}\mathbf{s}^t - \mathbf{x}^t||^2 + \sum_{j} |s_j^t|, \quad \text{for } t = 1, \ldots, T. \tag{5}$$

Or, in the absence of noise, by solving

$$\min_{\mathbf{s}^t} \sum_{j} |s_j^t| \quad \text{subject to} \quad \mathbf{A}\mathbf{s}^t = \mathbf{x}^t, \quad \text{for } t = 1, \ldots, T, \tag{6}$$

which can be formulated as a linear programming problem [4].

The mixing matrix $\mathbf{A}$ can be estimated either beforehand with a clustering algorithm, as shown in Section 4, or by external optimization

$$\min_{\mathbf{A}} \sum_{jt} |s_j^t(\mathbf{A})|, \tag{7}$$

where the $s_j^t(\mathbf{A})$'s represent, at each iteration, the solution of equation (6) using the current estimate of $\mathbf{A}$.

A similar two-step approach can be found in [8, 7], where a learning rule for $\mathbf{A}$ is derived by fitting a multivariate Gaussian around the current estimate of the source components.

3. SPARSITY AND SELECTION OF THE REPRESENTATION DOMAIN

Very often the raw data does not satisfy the requirement of sparsity. In the case of a square mixing matrix good results can still be obtained, but in the underdetermined case sparsity is a requirement for a good *separability* of the sources, even if the mixing matrix is known. In either case, when the data is *not* sparse enough, a useful approach is to use a reversible linear transform of the mixtures into a domain with improved sparsity, realize the separation in that domain, and transform the recovered sources back into the original domain. This approach was proposed in [12], and used in the square mixing matrix situation with successful results.

The benefits of such an approach are clear in Figure 1, where the FFT transform was used. Six flute signals playing different notes (see the SixFlutes example in Section 6) were synthetically mixed into two mixtures along equally spaced directions. Figure 1a presents a scatter plot of the resulting data (x_2^t against x_1^t for every t), showing a single big cloud. As it can be seen, the different sources are indistinguishable. Then each mixture was FFT-transformed and the scatter plot of the data in the frequency domain is shown in Figure 1b (i.e., x_2^w against x_1^w for every w). The difference is extraordinary. Now almost all the data points are neatly clustered along the six directions of the columns of the mixing matrix, thus providing very good separability.

In an abuse of notation, the temporal index t is extended hereafter to any representation domain (i.e., time samples, frequency bins, etc.), and the actual working domain will be made explicit when necessary.

4. POTENTIAL-FUNCTION BASED CLUSTERING FOR ESTIMATING THE MIXING MATRIX

As has just been shown, when the signals are sparse the distribution of data in mixture space forms a set of elongated clusters along the directions of the columns of the mixing matrix. Estimating the mixing matrix, then, consists of finding the directions of maximum data density. Our approach consists of defining a potential function in the space of directions. We associate a local potential function with direction of each data point, and then compute the potential function as the sum of the individual contributions. Local maxima of the potential function correspond to the mixing directions of the sources.

This approach is developed here for the case $M = 2$,

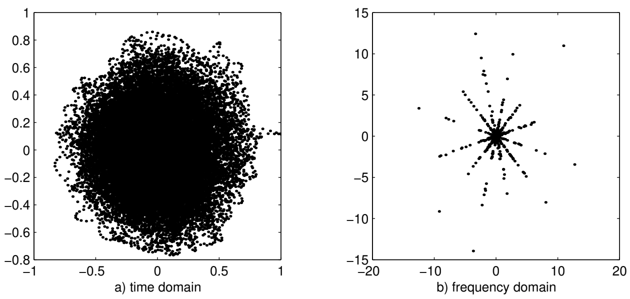

Figure 1: Scatter plot $\mathbf{X}_{2\cdot}$ vs $\mathbf{X}_{1\cdot}$ of six flute notes mixed into two mixtures along equally spaced directions in the (a) time and (b) frequency domains.

when mixture space is a plane and the directions can be parametrized using one dimension, an angle θ. Let $l_t = \sqrt{(x_1^t)^2 + (x_2^t)^2}$ and $\theta_t = \tan^{-1}(x_2^t/x_1^t)$ be the radius and angle, respectively, of data point $\mathbf{x}^t$, and let α be the relative angle between an arbitrary direction and θ_t. We define a *local potential function* ϕ_t around data point $\mathbf{x}^t$ as a triangular function of the angle α, scaled by the point radius l_t:

$$\begin{aligned} \phi_t(\alpha) &= l_t(1 - \frac{\alpha}{\pi/4}), \quad \text{for } |\alpha| < \pi/4, \\ &= 0 \quad \text{elsewhere.} \end{aligned} \tag{8}$$

We next span all possible directions (i.e, from 0 to π rad) with an equally spaced grid with K grid points $\theta_k = \pi/2K + k\pi/K$, $k = 1, \ldots, K$, and we compute the angles between each data point t and grid direction k, $\alpha_{tk} = |\theta^t - \theta_k|$. Then, we define the *potential function* as

$$\Phi(k, \lambda) = \sum_t \phi_t(\lambda\alpha_{tk}) \tag{9}$$

with λ a parameter to adjust the desired angular width or resolution of the local contributions. Spanning on a single dimension (the angle), local maxima of $\Phi(k)$ are easily detected and correspond to the directions of the columns of the mixing matrix. Notice that with this approach it is not necessary to know the *number* of sources beforehand, since it is *estimated* by the number of local maxima in the potential function.

The computational cost of this algorithm is $\mathcal{O}(T \times K)$. In practice, and without loss of performance, the number of data points T can be reduced significantly by discarding those with smallest relevance. That is, by running the clustering algorithm only with the data satisfying $l_t > h$, with h an adjustable threshold. On the other hand, too small a number K of grid points is undesirable, because the sampling resolution $(180/2K)$ would be too poor. Yet, when high accuracy is required, one can start the computations with a coarser grid, and refine the results later with a thinner grid in the neighborhood of the obtained cluster centers. Further improvement of accuracy may be obtained by continuous maximization of the potential function (9) for each cluster independently, or with the optimization strategy of Section 2.

A similar approach to the estimation of the mixing matrix was described in [11] for the case $M = N = 2$, using a histogram rather than a potential function, and image analysis tools based on self organizing maps were used in [9] for the analysis of several input distributions, including ICA.

5. A SHORTEST PATH DECOMPOSITION FOR THE SOURCES

Even when the mixing matrix is known, since the system in equation (1) is underdetermined, its solution is not unique. The sparse approach to ICA consists of finding the solution that minimizes the l_1 norm, as in equation (6). When the columns $\mathbf{a}^j$ are normalized, the optimal representation of the data point

$$\mathbf{x}^t = \sum_j \mathbf{a}^j s_j^t$$

that minimizes $\sum_j |s_j|$, will include at most N of the $\mathbf{a}^j$'s, corresponding to the vertices of the *minimal simplex* enclosing the direction of vector $\mathbf{x}^t$. The non-zero components of the optimal decomposition correspond then to the *shortest path* from the origin to the data point, when only the directions of the mixing matrix may be used.

In particular, for the two-sensor case (see Figure 2), the shortest path is obtained by choosing the columns $\mathbf{a}^b$ and $\mathbf{a}^a$ whose directions $\tan^{-1}(a_2^b/a_1^b)$ and $\tan^{-1}(a_2^a/a_1^a)$ are the closest from below and from above, respectively, to the direction θ^t.

Let $\mathbf{W}_r = [\mathbf{a}^b\mathbf{a}^a]^{-1}$ be the *reduced* $N \times N$ inverse matrix, and let $\mathbf{s}_r^t$ be the *reduced* decomposition along directions $\mathbf{a}^b$ and $\mathbf{a}^a$. The components of the sources are then obtained as

$$\begin{aligned} \mathbf{s}_r^t &= \mathbf{W}_r\mathbf{x}^t, \\ s_j^t &= 0, \quad \text{for} \quad j \neq b, a. \end{aligned} \tag{10}$$

In practice, $\mathbf{W}_r$ need only be computed once for all data points between any two pairs of mixing directions.

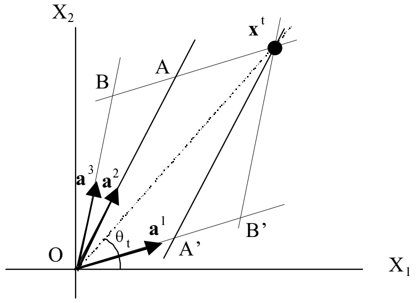

Figure 2: The *shortest path* from the origin to the data point $\mathbf{x}^t$ constrained by the directions $\mathbf{a}^j$ of the mixing matrix is O-A-$\mathbf{x}^t$ (or O-A'-$\mathbf{x}^t$). Therefore, $\mathbf{x}^t$ decomposes as O-A' along direction $\mathbf{a}^1$, as O-A along direction $\mathbf{a}^2$ and zero along direction $\mathbf{a}^3$. Notice that $\mathbf{x}^t$ has components different from zero only along those directions that *enclose* θ_t from below and above (i.e., $\mathbf{a}^1$ and $\mathbf{a}^2$, respectively).

6. EXPERIMENTS AND RESULTS

The approach was first tested using the SixFlutes data set: the sound of a flute playing steady, isolated notes was recorded at high-quality in an acoustically isolated booth without reverberation, and sampled at 44.1Khz with 16 bits resolution [1]. Six 743 ms excerpts (32768 samples) were selected for the sources, corresponding to the notes a4, d5, f5, g5, c6 and d#6. These six sources were then normalized so that they all had the same energy level and mixed into two mixtures along equally spaced directions. Without altering the relative levels, the dynamic range of the mixtures was adjusted to fit in the (-1,1) interval. Each of the mixture signals was then processed with a 32768 sample FFT (i.e., the whole length of the excerpts) and the real and imaginary parts of the positive spectra were used as input to the separation system. In experiment SixFlutes I, the clustering algorithm of Section 4 was run with parameter $\lambda = 5$ and threshold $h = 0.3$, and a grid with $K = 30$ equally spaced directions was used, so that each of the matrix directions was perfectly aligned with one of the grid points. The clustering was successful and the mixing matrix was precisely recovered. Results are shown in Table 1. When listening to the reconstructed signals the correct notes were very clear, but a little background noise was present (the accumulated sounds of the player blowing into the flute, plus some traces of cross-talk). Similar results were obtained with the external optimization approach of Section 2, when the starting state was not too far from the solution, or it would get trapped in local minima.

Table 1: S/N reconstruction indices (dB) for the different experiments (see text).

SixFlutes I	50.5	52.5	49.4	43.4	49.1	51.8
SixFlutes II	41.2	36.0	50.8	41.7	35.6	42.5
SixFlutes III	47.7	42.8	43.3	37.2	47.2	54.0
SixFlutes IV	-1.9	-2.0	-2.2	-2.4	-2.3	-2.4
FourVoices	21.7	19.4	15.7	16.6		
FiveSongs	15.6	15.5	15.0	15.1	15.2	
SixFlutMelod	20.4	19.4	14.2	16.1	24.7	29.1

The experiment was then repeated several times using random mixing matrices. The matrix was always correctly estimated within the 3 degrees of accuracy provided by the grid, but the reconstruction indices dropped. The following two experiments were then devised to measure the sensitivity of the separation algorithm to the accuracy in the estimation of the matrix and to the closeness of the sources, independently. Experiment SixFlutes II was identical to SixFlutes I except for the grid points, which were shifted 3 degrees from their original positions and therefore were no longer aligned with the mixing directions. After clustering, the estimated mixing directions were all off by 3 degrees, as expected, and results of the separation (Table 1) were impaired by 8.1dB in average . In experiment SixFlutes III the mixing directions were lumped together in a total span of 6 degrees, so that each source direction was separated by only one degree from the next. The number of grid points was set to the $K = 540$ so as to guarantee the alignment, and $\lambda = 55$ was required to get enough resolution. With this setting, the mixing matrix was again perfectly recovered and separation indices are shown in Table 1. The loss was now only 4.1dB in average with respect to the SixFlutes I experiment, which illustrates the relative insensitivity of the separation procedure to the proximity between the sources.

For the sake of comparison, the last experiment (SixFlutes IV) was conducted on the same data set using the mixtures in the time domain instead of in the frequency domain. The maxima of the potential function were no longer in the directions of the mixing matrix, so the resulting estimate was meaningless. The separation was then attempted using the original mixing matrix, but the algorithm totally failed to separate the sources, as shown in Table 1.

The flute notes in the SixFlutes data set above were very steady, which allowed for a very large FFT window size. The remaining three experiments presented here were performed on much more dynamic signals, and preprocessing was required based on a short-term, frame-by-frame analysis. As before, the sources were first normalized to the same energy level[1], mixed in the time domain, and the mixtures rescaled so as to fit in the (-1,1) interval. Each mixture was then processed with a Hanning window of length L, and a "hop" distance d between the starting point of successive frames (i.e, overlapping $L-d$). Each frame was FFT-transformed and the real and imaginary parts of the positive spectra were obtained. For each mixture, the input to the separation system was then a single vector containing the concatenation of the coefficients of all the frames in that mixture. After the separation (post-processing), the estimated signals were resynthesized by reconstructing the frames, regrouping the real and imaginary parts, taking inverse FFT and inverse windowing. The overlap was removed by keeping only the central part of the frame (thus avoiding the distortion at the edges that often appears after frequency domain manipulation) and the reconstructed signal was obtained by simple concatenation of the resulting pieces.

The experiments were conducted on the following sets of signals: A FourVoices data set with four 2.9 sec sentences pronounced by four different people (three females and a male), recorded at 22,050 Hz and 8 bits with a low quality microphone on a home personal computer. Pre-processing was done with $L = 2048$ and $d = 614$ samples. A FiveSongs data set with five 5 sec long full-ensemble music pieces (two classical and three pop/folk music) extracted from standard CDs (44,100 Hz/16 bits), downsampled to 11,025 Hz monophonic and processed with $L = 4096$ and $d = 1228$ samples. Finally, a SixFluteMelodies data set [1] including six 5.7 sec long flute melodies (the two voices of a canon, the two voices of a duet and two unrelated melodies) with a high-quality registration at 44,100 Hz/16 bits, down-sampled to 22,050 Hz and processed with $L = 8192$ and $d = 3276$ samples.

In all three cases the mixing matrix was formed with equally spaced directions, and the number of grid points was selected for perfect alignment ($K = 36, 35$ and 30, respectively) in order to be able to measure the maximum separation ability of the system. As the SixFlutes experiments had shown, the estimation of the mixing matrix was always successful, showing very little sensitivity to λ (equally good results were obtained with λ in the range from 3 to 50). Results of the separation for $\lambda = 5$ and $h = 0.3$ are shown in Table 1. Although good enough in themselves, the reconstruction indices of the dynamic signals were significantly poorer than those of the SixFlutes I experiment, in part due to the intrinsic difficulties of the short-term analysis and resynthesis. Reconstruction indices were on the same range for the three examples, regardless of the number of voices, with somehow worse results in the case of the FiveSongs, probably due to the higher complexity of the sounds. The plot of the recovered signals was in all cases very similar to the plot of the original sources, as illustrated in Figure 3 for the FourVoices case. From a subjective listening point of view, the separation of the FourVoices example was remarkable for the high intelligibility of the recovered sentences, in spite of some background noise and cross-talk. In the case of the FiveSongs, the reconstructed songs were also very clear but the quality of the sound was sensibly degraded by background noise, cross-talk and a flattening of percussive sounds and sharp transitions. Finally, in the SixFluteMelodies example, although the recovered melodies were clear, a sort of ringing artifact appeared in the transitions between notes, and some frame-rate rattling noise was present. Sound examples for the above experiments are available on-line in [3].

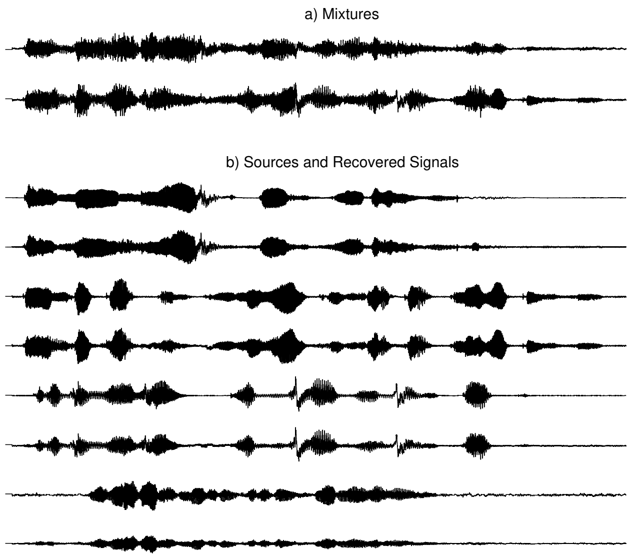

Figure 3: FourVoices experiment. (a) Mixtures. (b) Sources and recovered signals (pairwise).

7. DISCUSSION AND FURTHER WORK

In the context of underdetermined Blind Source Separation (i.e., BSS with fewer mixtures than sources), the three main contributions of this paper have been the

[1] In fact, energy normalization of the sources was only required for the FourVoices data set, described latter, since one of the voices had been recorded at a much lower level than the rest. In the other cases, without significantly affecting their quality, the results it yielded were more balanced.

benefits of performing blind source separation in the frequency domain (rather than in the time domain); a clustering algorithm for the estimation of the mixing matrix in the two-sensor case; and a shortest path separation procedure that yields the most sparse reconstruction of the sources from two mixtures. Several experiments have been presented involving music and speech signals, with rather good results, including the successful separation of six sources from two mixtures.

From the above three points the most effective contribution to a successful separation has probably been the exploitation of sparsity in the frequency domain since, as experiments have shown, only the transformed data satisfy the assumptions of sparsity required by the clustering and separation algorithms. The estimation of the mixing matrix has always been successful, within the accuracy provided by the sampling grid, and the separation was more adversely affected by an inaccurate estimate of the mixing directions than by the proximity of the mixed sources to each other. S/N reconstruction indices have shown excellent scores for steady flute notes that could be processed with a single FFT, and good scores for the other three dynamic examples, which required short-term analysis and resynthesis. The recovered signals have been highly intelligible to the ear in all cases, in spite of some background noise and some cross-talk. Results seem to show that the difficulty of the separation depends more on the complexity of the sounds than on the number of sources present, but further experiments would be required in order to assess this trend.

Even if l_1 norm minimization is theoretically the most likely *a posteriori* estimation for Laplacian sources, in practice good separation is obtained only where the sources are *disjoint* or almost disjoint, regardless of whether they are Laplacian. This is usually the case for the overtones of signals with different pitch, for instance. But when sources overlap, the shortest path separation criterion, although statistically optimal, is unable to give the credit to the sources actually involved.

As further work, we are currently interested in the extension of the methods presented here to any number of mixtures, the improvement of the pre- and post-processing steps, and the evaluation of other transforms (Gabor, wavelet, etc) that might be better adapted to transitions, or lead to improved sparsity (and hopefully disjointness) of the sources.

8. REFERENCES

[1] All flute examples performed by Linda Antas, University of Washington.

[2] Bell A.J. and Sejnowski T.J., "An Information-Maximization Approach to Blind Separation and Blind Deconvolution", in *Neural Computation*, Vol 7 (6), pp 1129-1159, 1995. http://www.cnl.salk.edu/cgi-bin/pub-search#articles

[3] Bofill P. and Zibulevsky M., *Sound Examples*, at http://www.ac.upc.es/homes/pau/

[4] Chen S.S., Donoho D.L. and Saunders A., "Atomic Decomposition by Basis Pursuit", TR, http://www-stat.stanford.edu/~donoho/Reports/

[5] Hyvärinen A., "Survey on Independent Component Analysis", in *Neural Computing Surveys*, No 2, pp 94-128, 1999 http://www.cis.hut.fi/projects/ica/

[6] Jutten C. and Herault J., "Blind Separation of Sources, an Adaptive Algorithm Based on Neuromimetic Architecture", in *Signal Processing*, Vol 24, No 1, pp 1-10, 1991.

[7] Lee T-W., Lewicki M.S., Girolami M. and Sejnowski T.J., "Blind Source Separation of More Sources than Mixtures Using Overcomplete Representations", in *IEEE Signal Processing Letters*, Vol 6, No 4, pp 87-90, 1999.

[8] Lewicki M.S. and Sejnowski T.J., "Learning Overcomplete Representations", in *Neural Computation*, in press, 1998. http://www.cs.cmu.edu/~lewicki

[9] Lin J.K. and Grier G., "Faithful Representation of Separable Distributions", in *Neural Computation*, Vol 9, pp 1305-1320, 1997.

[10] Olshausen B.A. and Field D.J., "Sparse coding with an overcomplete basis set: A strategy employed by V1?", in *Vision Research*, 37: 3311-3325, 1997.

[11] Prieto A., Prieto B., Puntonet C.G., Cañas A. and Martín-Smith P., "Geometric Separation of Linear Mixtures of Sources: Application to Speech Signals", in *Proc. of 1st Int. Workshop on ICA and Signal Sep. (ICA'99)*, pp 295-300, January 1999. http://atc.ugr.es/~bprieto/sfuentes/articulo.html

[12] Zibulevsky M. and Pearlmutter B.A., "Blind Source Separation by Sparse Decomposition", TR No. CS99-1, University of New Mexico, Albuquerque, July 1999. http://www.cs.unm.edu/~bap/papers/sparse-ica-99a.ps.gz