Manuscript NC 2079

Blind Source Separation by Sparse Decomposition in a Signal Dictionary

Michael Zibulevsky
Dept. of Computer Science
University of New Mexico
Albuquerque, NM 87131

Barak A. Pearlmutter
Dept. of Computer Science
Dept. of Neurosciences
University of New Mexico
Albuquerque, NM 87131

July 9, 2000
(Final *Neural Computation* copy)

Abstract

The blind source separation problem is to extract the underlying source signals from a set of linear mixtures, where the mixing matrix is unknown. This situation is common, in acoustics, radio, medical signal and image processing, hyperspectral imaging, *etc.*. We suggest a two-stage separation process. First, *a priori* selection of a possibly overcomplete signal dictionary (for instance a wavelet frame, or a learned dictionary) in which the sources are assumed to be sparsely representable. Second, unmixing the sources by exploiting the their sparse representability. We consider the general case of more sources than mixtures, but also derive a more efficient algorithm in the case of a non-overcomplete dictionary and an equal numbers of sources and mixtures. Experiments with artificial signals and with musical sounds demonstrate significantly better separation than other known techniques.

1 Introduction

In blind source separation an N-channel sensor signal $x(t)$ arises from M unknown scalar source signals $s_i(t)$, linearly mixed together by an unknown $N \times M$ matrix A, and possibly corrupted by additive noise $\xi(t)$

$$x(t) = As(t) + \xi(t) \tag{1}$$

We wish to estimate the mixing matrix A and the M-dimensional source signal $s(t)$. Many natural signals can be sparsely represented in a proper signal dictionary

$$s_i(t) = \sum_{k=1}^{K} C_{ik}\, \varphi_k(t) \tag{2}$$

The scalar functions $\varphi_k(t)$ are called *atoms* or *elements* of the dictionary. These elements do not have to be linearly independent, and instead may form an overcomplete dictionary. Important examples are wavelet-related dictionaries (wavelet packets, stationary wavelets, etc, see for example Chen et al. (1996); Mallat (1998) and references therein), or learned dictionaries (Lewicki and Sejnowski, 1998; Lewicki and Olshausen, 1999; Olshausen and Field, 1997, 1996). Sparsity means that only a small number of the coefficients C_{ik} differ significantly from zero.

We suggest a two stage separation process. First, *a priori* selection of a possibly overcomplete signal dictionary in which the sources are assumed to be sparsely representable. Second, unmixing the sources by exploiting their sparse representability.

In the discrete time case $t = 1, 2, \dots, T$ we use matrix notation. X is an $N \times T$ matrix, with the i-th component $x_i(t)$ of the sensor signal in row i, S is an $M \times T$ matrix with the signal $s_j(t)$ in row j, and Φ is a $K \times T$ matrix with basis function $\varphi_k(t)$ in row k. Equations (1) and (2) then take the following simple form

$$X = AS + \xi \tag{3}$$

$$S = C\Phi \tag{4}$$

Combining them, we get the following when the noise is small

$$X \approx AC\Phi$$

Our goal therefore can be formulated as follows:

> *Given the sensor signal matrix X and the dictionary Φ, find a mixing matrix A and matrix of coefficients C such that $X \approx AC\Phi$ and C is as sparse as possible.*

We should mention other problems of sparse representation studied in the literature. The basic problem is to represent sparsely scalar signal in given dictionary (see for example Chen et al. (1996) and references therein). Another problem is to adapt the dictionary to the given class of signals[1] (Lewicki and Sejnowski, 1998; Lewicki and Olshausen, 1999; Olshausen

[1] Our dictionary Φ may be obtained in this way.

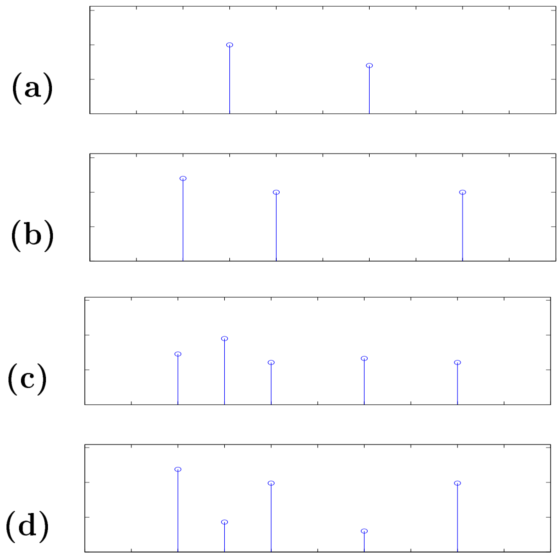

Figure 1: Sources (a and b) are sparse. Mixtures (c and d) are less sparse.

and Field, 1997). This problem is shown to be equivalent to the problem of blind source separation, when the sources are sparse in time (Lee et al., 1998; Lewicki and Sejnowski, 1998). Our problem is different, but we will use and generalize some techniques presented in these works.

Independent Factor Analysis (Attias, 1999) and the Bayesian blind source separation (Rowe, 1999) also consider the case of more sources than mixtures. In our approach we take an advantage, when the sources are sparsely representable. In extreme case, when the decomposition coefficients are very sparse, the separation becomes practically ideal (see Section 3.2 below, and the six flutes example in Zibulevsky et al. (2000)). Nevertheless detailed comparison of the methods on real-world signals remains open for future research.

Our paper is organized as follows. In Section 2 we give some motivating examples, which demonstrate how sparsity helps to separate sources. Section 3 gives the problem formulation in probabilistic framework, and presents the *maximum a posteriori* approach, which is applicable to the case of more sources than mixtures. In Section 4 we derive another objective function, which provides more robust computations when there are an equal number of sources and mixtures. Section 5 presents sequential source extraction using quadratic programming with non-convex quadratic constraints. Finally, in Section 6 we derive a faster method for non-overcomplete dictionaries and demonstrate high-quality separation of synthetically mixed musical sounds.

2 Separation of Sparse Signals

In this section we present two examples which demonstrate how sparsity of source signals in the time domain helps to separate them. Many real-world signals have sparse representations in a proper signal dictionary, but not in the time domain. The intuition here carries over to that situation, as shown in Section 3.1.

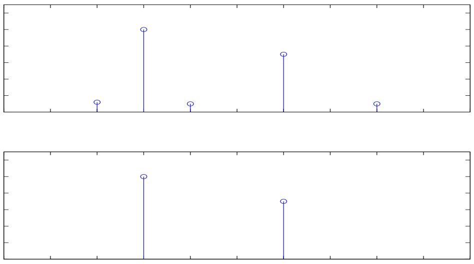

Figure 2: (a) Imperfect separation. Since the second source is not completely removed, the total number of non-zero samples remains five. (b) Perfect separation. When the source is recovered perfectly, the number of non-zero samples drops to two and the objective function achieves its minimum.

Example: 2 sources and 2 mixtures. Two synthetic sources are shown in Figure 1(a,b). The first source has two non-zero samples, and the second has three. The mixtures, shown in Figure 1(c,d) are less sparse: they have five non-zero samples each. One can use this observation to recover the sources. For example, we can express one of the sources as

$$\tilde{s}_i(t) = x_1(t) + \mu x_2(t)$$

and chose μ such as to minimize the number of non-zero samples $\|\tilde{s}_i\|_0$, *i.e.* the l_0 norm of s_i.

This objective function yields perfect separation. As shown in Figure 2(a), when μ is not optimal the second source interferes, and the total number of non-zero samples remains five. Only when the first source is recovered perfectly, as in Figure 2(b), does the number of non-zero samples drop to two, and the objective function achieve its minimum.

Note that the function $\|\tilde{s}_i\|_0$ is discontinuous and may be difficult to optimize. It is also very sensitive to noise: even a tiny bit of noise would make all the samples non-zero. Fortunately in many cases the l_1 norm $\|\tilde{s}_i\|_1$ is a good substitute for this objective function. In this example, it too yields perfect separation.

Example: 3 sources and 2 mixtures. The signals are presented in Figure 3.

These sources have about 10% non-zero samples. The non-zero samples have random positions, and are zero-mean unit-variance Gaussian distributed in amplitude. Figure 4 shows a scatter plot of the mixtures. The directions of the columns of mixing matrix are clearly visible. This phenomena can be used in clustering approaches to source separation (Pajunen et al., 1996; Zibulevsky et al., 2000). In this work we will explore a maximum *a posteriori* approach.

3 Probabilistic Framework

In order to derive a maximum *a posteriori* solution, we consider the blind source separation problem in a probabilistic framework (Belouchrani and Cardoso, 1995; Pearlmutter and Parra, 1996). Suppose that the coefficients C_{ik} in a source decomposition (4) are independent random variables with a probability density function (pdf) of an exponential type

$$p_i(C_{ik}) \propto \exp -\beta_i h(C_{ik}) \tag{5}$$

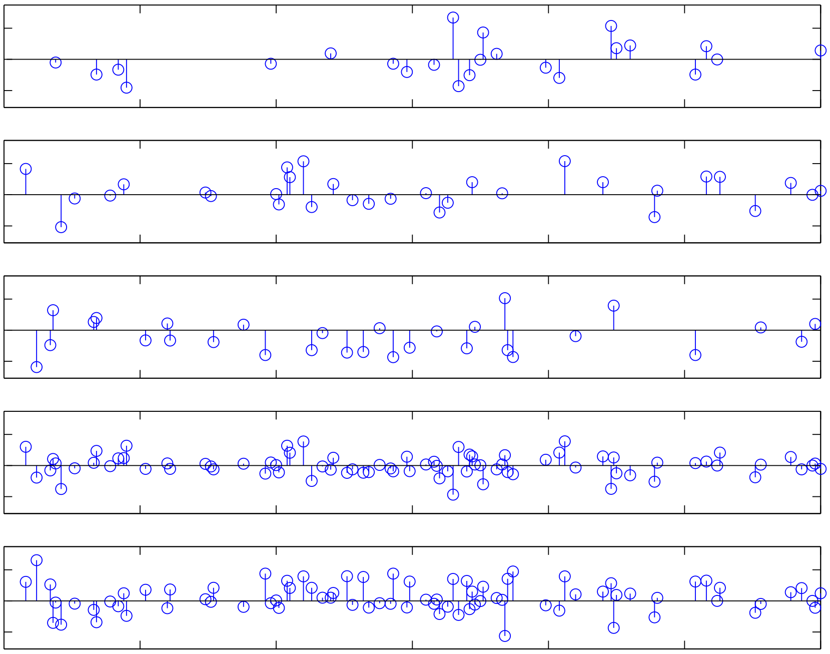

Figure 3: Top three panels: sparse sources (sparsity is 10%). Bottom two panels: mixtures.

This kind of distribution is widely used for modeling sparsity (Lewicki and Sejnowski, 1998; Olshausen and Field, 1997). A reasonable choice of $h(c)$ may be

$$h(c) = |c|^{1/\gamma} \qquad\qquad \gamma \geq 1 \tag{6}$$

or a smooth approximation thereof. Here we will use a family of convex smooth approximations to the absolute value

$$h_1(c) \;=\; |c| - \log(1 + |c|) \tag{7}$$
$$h_\lambda(c) \;=\; \lambda h_1(c/\lambda) \tag{8}$$

with λ a proximity parameter: $h_\lambda(c) \to |c|$ as $\lambda \to 0^+$.

We also suppose *a priori* that the mixing matrix A is uniformly distributed over the range of interest, and that the noise $\xi(t)$ in (3) is a spatially and temporally uncorrelated Gaussian process[2] with zero mean and variance σ^2.

3.1 Maximum a posteriori approach

We wish to maximize the posterior probability

$$\max_{A,C} P(A, C|X) \;\propto\; \max_{A,C} P(X|A, C)\, P(A)\, P(C) \tag{9}$$

[2]The assumption that the noise is white is for simplicity of exposition, and can be easily removed.

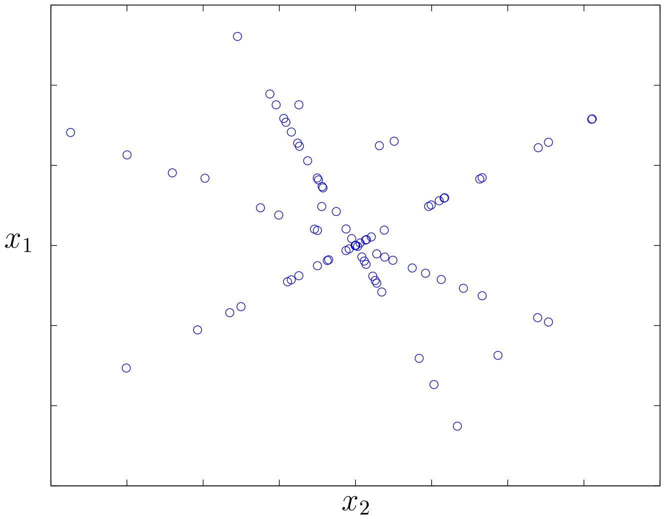

Figure 4: Scatter plot of two sensors. Three distinguished directions, which correspond to the columns of the mixing matrix A, are visible.

where $P(X|A,C)$ is the conditional probability of observing X given A and C. Taking into account (3), (4), and the white Gaussian noise, we have

$$P(X|A,C) \propto \prod_{i,t} \exp -\frac{(X_{it} - (AC\Phi)_{it})^2}{2\sigma^2} \tag{10}$$

By the independence of the coefficients C_{jk} and (5), the prior pdf of C is

$$P(C) \propto \prod_{j,k} \exp(-\beta_j h(C_{jk})) \tag{11}$$

If the prior pdf $P(A)$ is uniform, it can be dropped[3] from (9). In this way we are left with the problem

$$\max_{A,C} P(X|A,C)\, P(C). \tag{12}$$

By substituting (10) and (11) into (12), taking the logarithm, and inverting the sign, we obtain the following optimization problem

$$\min_{A,C} \frac{1}{2\sigma^2} \|AC\Phi - X\|_F^2 + \sum_{j,k} \beta_j h(C_{jk}) \tag{13}$$

where $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2}$ is the Frobenius matrix norm.

One can consider this objective as a generalization of Olshausen and Field (1996, 1997) by incorporating the matrix Φ, or as a generalization of Chen et al. (1996) by including the matrix A. One problem with such a formulation is that it can lead to the degenerate solution

[3]Otherwise, if $P(A)$ is some other known function, we should use (9) directly.

$C = 0$ and $A = \infty$. We can overcome this difficulty in various ways. The first approach is to force each row A_i of the mixing matrix A to be bounded in norm,

$$\|A_i\| \leq 1 \qquad\qquad i = 1, \ldots, N. \tag{14}$$

The second way is to restrict the norm of the rows C_j from below

$$\|C_j\| \geq 1 \qquad\qquad j = 1, \ldots, M. \tag{15}$$

A third way is to reestimate the parameters β_j based on the current values of C_j. For example, this can be done using sample variance as follows: for a given function $h(\cdot)$ in the distribution (5), express the variance of C_{jk} as a function $f_h(\beta)$. An estimate of β can be obtained by applying the corresponding inverse function to the sample variance,

$$\hat{\beta}_j = f_h^{-1}(K^{-1} \sum_k C_{jk}^2) \tag{16}$$

In particular, when $h(c) = |c|$, $\mathrm{var}(c) = 2\beta^{-2}$ and

$$\hat{\beta}_j = \frac{2}{\sqrt{K^{-1} \sum_k C_{jk}^2}} \tag{17}$$

Substituting $h(\cdot)$ and $\hat{\beta}$ into (13), we obtain

$$\min_{A,C} \frac{1}{2\sigma^2} \|AC\Phi - X\|_F^2 + \sum_j \frac{2 \sum_k |C_{jk}|}{\sqrt{K^{-1} \sum_k C_{jk}^2}} \tag{18}$$

This objective function is invariant to a rescaling of the rows of C combined with a corresponding inverse rescaling of the columns of A.

3.2 Experiment: more sources than mixtures

This experiment demonstrates that sources which have very sparse representations can be separated almost perfectly, even when they are correlated and the number of samples is small.

We used the standard wavelet packet dictionary with the basic wavelet *symmlet-8*. When the signal length is 64 samples, this dictionary consists of 448 atoms *i.e.* it is overcomplete by a factor of seven. Examples of atoms and their images in the time-frequency phase plane (Coifman and Wickerhauser, 1992; Mallat, 1998) are shown in Figure 5. We used the ATOMIZER (Chen et al., 1995) and WAVELAB (Buckheit et al., 1995) MATLAB packages for fast multiplication by Φ and Φ^T.

We created three very sparse sources (Figure 6(a)), each composed of only two or three atoms. The first two sources have significant cross-correlation, equal to 0.34, which makes separation difficult for conventional methods. Two synthetic sensor signals (Figure 6(b)) were obtained as linear mixtures of the sources. In order to measure the accuracy of separation, we

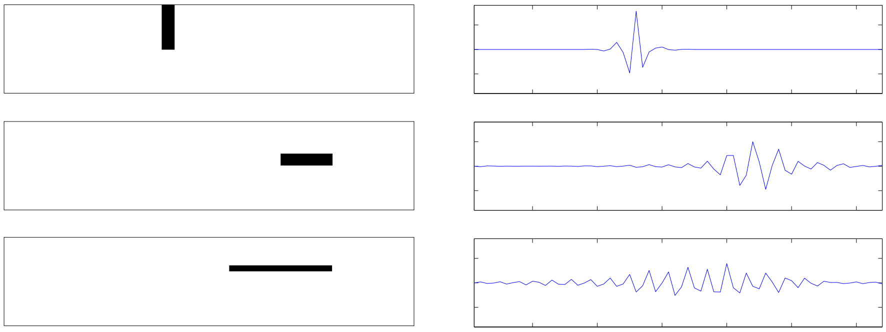

Figure 5: Examples of atoms: time-frequency phase plane (left) and time plot (right.)

normalized the original sources with $\|S_j\|_2 = 1$, and the estimated sources with $\|\tilde{S}_j\|_2 = 1$. The error was computed as

$$\text{Error} = \frac{\|\tilde{S}_j - S_j\|_2}{\|S_j\|_2} \cdot 100\% \tag{19}$$

We tested two methods with this data. The first method used the objective function (13) and the constraints (15), while the second method used the objective function (18). We used PBM (Ben-Tal and Zibulevsky, 1997) for the constrained optimization. The unconstrained optimization was done using the method of conjugate gradients, with the TOMLAB package (Holmstrom and Bjorkman, 1999). The same tool was used by PBM for its internal unconstrained optimization.

We used $h_\lambda(\cdot)$ defined by (7) and (8) with $\lambda = 0.01$ and $\sigma^2 = 0.0001$ in the objective function. The resulting errors of the recovered sources were 0.09% and 0.02% by the first and the second methods, respectively. The estimated sources are shown in Figure 6(c). They are visually indistinguishable from the original sources in Figure 6(a).

It is important to recognize the computational difficulties of this approach. First, the objective functions seem to have multiple local minima. For this reason, reliable convergence was achieved only when the search started randomly within 10%–20% distance to the actual solution (in order to get such an initial guess one can use a clustering algorithm, as in Pajunen et al. (1996) or Zibulevsky et al. (2000).)

Second, the method of conjugate gradients requires a few thousand iterations to converge, which takes about 5 min on a 300 MHz AMD K6-II even for this very small problem. (On the other hand, preliminary experiments with a truncated Newton method have been encouraging, and we anticipate that this will reduce the computational burden by an order of magnitude or more. Also Paul Tseng's block coordinate descent method (unpublished manuscript) may be appropriate.) Below we present a few other approaches which help to stabilize and accelerate the optimization.

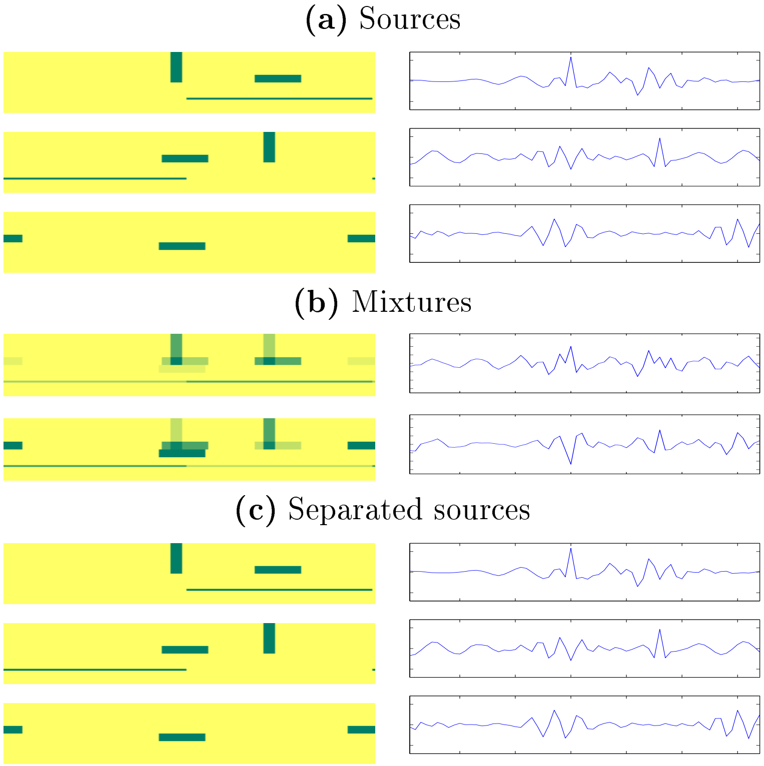

Figure 6: Sources, mixtures and reconstructed sources, in both time-frequency phase plane (left) and time domain (right).

4 Equal number of sources and sensors: more robust formulations

The main difficulty in a maximization problem like (13) is the bilinear term $AC\Phi$, which destroys the convexity of the objective function and makes convergence unstable when optimization starts far from the solution. In this section we consider more robust formulations for the case when the number of sensors is equal to the number of sources, $N = M$, and the mixing matrix is invertible, $W = A^{-1}$.

When the noise is small and the matrix A is far from singular, WX gives a reasonable estimate of the source signals S. Taking into account (4), we obtain a least squares term $\|C\Phi - WX\|_F^2$, so the separation objective may be written

$$\min_{W,C} \frac{1}{2}\|C\Phi - WX\|_F^2 + \mu \sum_{j,k} \beta_j h(C_{jk}) \tag{20}$$

We also need to add a constraint which enforces the non-singularity of W. For example, we can restrict its minimal singular value $r_{\min}(W)$ from below,

$$r_{\min}(W) \geq 1 \tag{21}$$

It can be shown that in the noiseless case, $\sigma \approx 0$, the problem (20)–(21) is equivalent to the maximum *a posteriori* formulation (13) with the constraint $\|A\|_2 \leq 1$. Another possibility

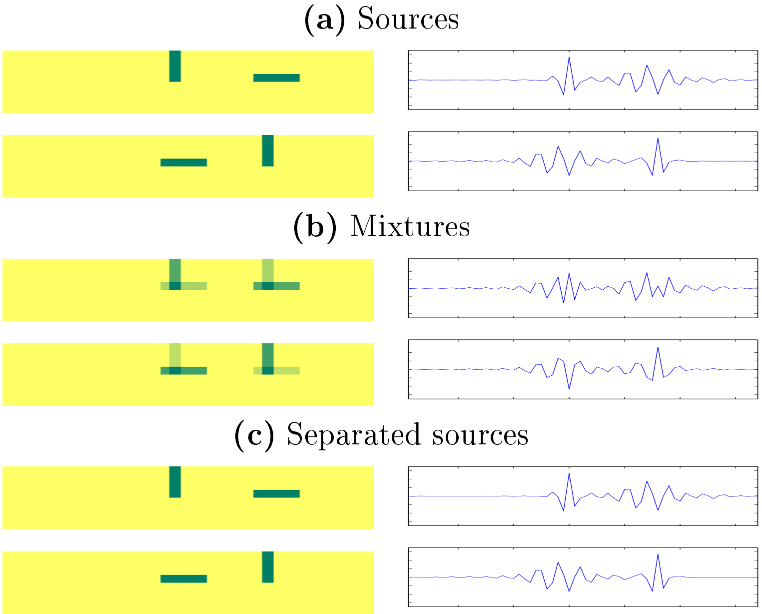

Figure 7: Sources, mixtures and reconstructed sources, in both time-frequency phase plane (left) and time domain (right).

for ensuring the non-singularity of W is to subtract $K \log|\det W|$ from the objective

$$\min_{W,C} \; -K \log|\det W| + \frac{1}{2}\|C\Phi - WX\|_F^2 + \mu \sum_{j,k} \beta_j h(C_{jk}) \tag{22}$$

which (Bell and Sejnowski, 1995; Pearlmutter and Parra, 1996) can be viewed as a maximum likelihood term.

When the noise is zero and Φ is the identity matrix, we can substitute $C = WX$ and obtain the BS Infomax objective (Bell and Sejnowski, 1995)

$$\min_{W} \; -K \log|\det W| + \sum_{j,k} \beta_j h((WX)_{jk}) \tag{23}$$

Experiment: equal numbers of sources and sensors. We created two sparse sources (Figure 7, top) with strong cross-correlation of 0.52. Separation by minimization of the objective function (22) gave an error of 0.23%. Robust convergense was achieved when we started from random uniformly distributed points in C and W.

For comparison we tested the JADE (Cardoso, 1999a), FastICA (Hyvärinen, 1999) and BS Infomax (Bell and Sejnowski, 1995; Amari et al., 1996) algorithms on the same signals. All three codes were obtained from public web sites (Cardoso, 1999b; Hyvärinen, 1998; Makeig, 1999) and were used with default setting of all parameters. The resulting relative errors (Figure 8) confirm the significant superiority of the sparse decomposition approach.

This still takes a few thousands conjugate gradient steps to converge (about 5 min on a 300 MHz AMD K6). For comparison, the tuned public implementations of JADE, FastICA and BS Infomax take only a few seconds. Below we consider some options for acceleration.

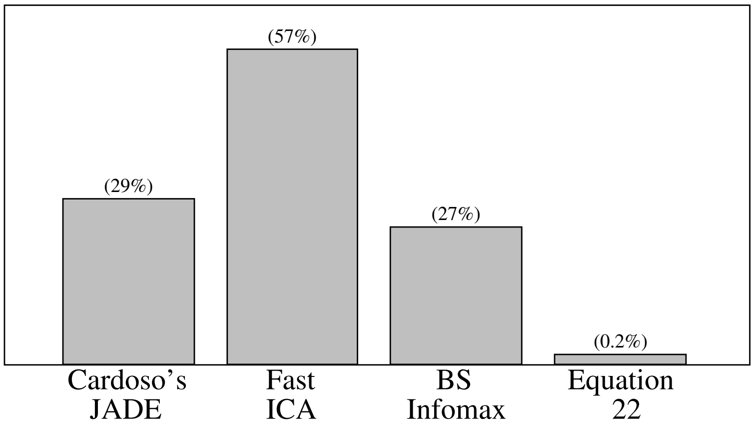

Figure 8: Percent relative error of separation of the artificial sparse sources recovered by (1) JADE, (2) Fast ICA, (3) Bell-Sejnowski Infomax, (4) Equation 22.

5 Sequential Extraction of Sources via Quadratic Programming

Let us consider finding the sparsest signal that can be obtained by a linear combination of the sensor signals $s = w^T X$. By sparsity we mean the ability of the signal to be approximated by a linear combination of a small number of dictionary elements φ_k, as $s \approx c^T \Phi$. This leads to the objective

$$\min_{w,c} \frac{1}{2} \|c^T \Phi - w^T X\|_2^2 + \mu \sum_k h(c_k), \tag{24}$$

where the term $\sum_k h(c_k)$ may be considered a penalty for non-sparsity. In order to avoid the trivial solution of $w = 0$ and $c = 0$ we need to add a constraint that separates w from zero. It could be, for example,

$$\|w\|_2^2 \geq 1 \ , \tag{25}$$

A similar constraint can be used as a tool to extract all the sources sequentially: the new separation vector w^j should have a component of unit norm in the subspace orthogonal to the previously extracted vectors $w^1, \ldots, w^{j-1}$

$$\|(I - P^{j-1}) w^j\|_2^2 \geq 1 \ , \tag{26}$$

where P^{j-1} is an orthogonal projector onto Span$\{w^1, \ldots, w^{j-1}\}$.

When $h(c_k) = |c_k|$ we can use the standard substitution

$$\begin{aligned} c &= c^+ - c^- \ , \quad c^+ \geq 0 \ , \quad c^- \geq 0 \\ \hat{c} &= \begin{pmatrix} c^+ \\ c^- \end{pmatrix} \quad \text{and} \quad \hat{\Phi} = \begin{pmatrix} \Phi \\ -\Phi \end{pmatrix} \end{aligned}$$

that transforms (24) and (26) into the quadratic program

$$\begin{aligned} \min_{w,\hat{c}} \quad & \frac{1}{2} \|\hat{c}^T \hat{\Phi} - w^T X\|_2^2 + \mu e^T \hat{c} \\ \text{subject to:} \quad & \|w\|_2^2 \geq 1 \ , \quad \hat{c} \geq 0 \end{aligned}$$

where e is a vector of ones.

6 Fast Solution in Non-overcomplete Dictionaries

In important applications (Tang et al., 1999, 2000a,b), the sensor signals may have hundreds of channels and hundreds of thousands of samples. This may make separation computationally difficult. Here we present an approach which compromises between statistical and computational efficiency. In our experience this approach provides high quality of separation in reasonable time.

Suppose that the dictionary is "complete," *i.e.* it forms a basis in the space of discrete signals. This means that the matrix Φ is square and non-singular. As examples of such a dictionary one can think of the Fourier basis, Gabor basis, various wavelet-related bases, *etc.* We can also obtain an "optimal" dictionary by learning from given family of signals (Lewicki and Sejnowski, 1998; Lewicki and Olshausen, 1999; Olshausen and Field, 1997, 1996).

Let us denote the dual basis

$$\Psi = \Phi^{-1} \tag{27}$$

and suppose that coefficients of decomposition of the sources

$$C = S\Psi \tag{28}$$

are sparse and independent. This assumption is reasonable for properly chosen dictionaries, although of course we would lose the advantages of overcompleteness.

Let Y be the decomposition of the sensor signals

$$Y = X\Psi \tag{29}$$

Multiplying both sides of (3) by Ψ from the right and taking into account (28) and (29), we obtain

$$Y = AC + \zeta \ , \tag{30}$$

where $\zeta = \xi\Psi$ is the decomposition of the noise. Here we consider an "easy" situation, where ζ is white, which assumes that Ψ is orthogonal. We can see that all the objective functions from the sections 3.1–5 remain valid if we substitute the identity matrix for Φ and replace the sensor signal X by its decomposition Y. For example, the maximum *a posteriori* objectives (13) and (18) are transformed into

$$\min_{A,C} \ \frac{1}{2\sigma^2}\|AC - Y\|_F^2 + \sum_{j,k} \beta_j h(C_{jk}) \tag{31}$$

and

$$\min_{A,C} \ \frac{1}{2\sigma^2}\|AC - Y\|_F^2 + \sum_{j} \frac{2\sum_k |C_{jk}|}{\sqrt{K^{-1}\sum_k C_{jk}^2}} \tag{32}$$

The objective (22) becomes

$$\min_{W,C} \ -K \log|\det W| + \frac{1}{2}\|C - WY\|_F^2 + \mu \sum_{j,k} \beta_j h(C_{jk}) \tag{33}$$

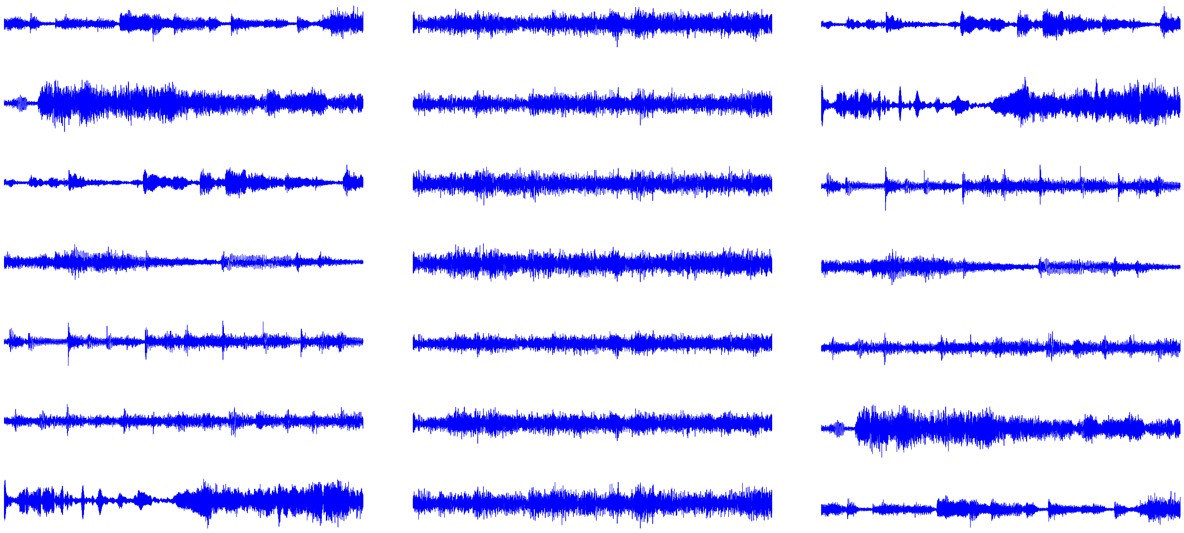

Figure 9: Separation of musical recordings taken from commercial digital audio CDs (five second fragments).

In this case we can further assume that the noise is zero, substitute $C = WY$, and obtain the BS Infomax objective (Bell and Sejnowski, 1995)

$$\min_{W} \; -K \log |\det W| + \sum_{j,k} \beta_j h((WY)_{jk}) \tag{34}$$

Also other known methods (for example, Lee et al. (1998); Lewicki and Sejnowski (1998)), which normally assume sparsity of source signals, may be directly applied to the decomposition Y of the sensor signals. This may be more efficient than the traditional approach, and the reason is obvious: typically, a properly chosen decomposition gives significantly higher sparsity for the transformed coefficients than for the raw signals. Furthermore, independence of the coefficients is a more realistic assumption than independence of the raw signal samples.

Experiment: musical sounds. In our experiments we artificially mixed seven 5-second fragments of musical sound recordings taken from commercial digital audio CDs. Each of them included 40k samples after down-sampling by a factor of 5. (Figure 9).

The easiest way to perform sparse decomposition of such sources is to compute a *spectrogram*, the coefficients of a time-windowed discrete Fourier transform. (We used the function SPECGRAM from the MATLAB signal processing toolbox with a time window of 1024 samples.) The sparsity of the spectrogram coefficients (the histogram in Figure 10, right) is much higher then the sparsity of the original signal (Figure 10, left)

In this case Y (29) is a real matrix, with separate entries for the real and imaginary components of each spectrogram coefficient of the sensor signals X. We used the objective function (34) with $\beta_j = 1$ and $h_\lambda(\cdot)$ defined by (7) and (8) with the parameter $\lambda = 10^{-4}$. Unconstrained minimization was performed by a BFGS Quasi-Newton algorithm (MATLAB function FMINU.)

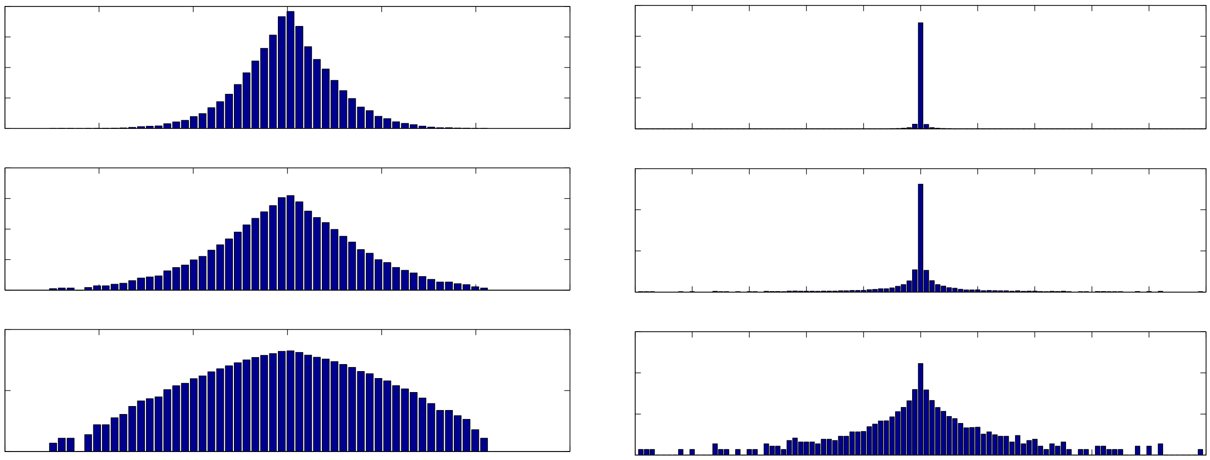

Figure 10: Histogram of sound source values (left) and spectrogram coefficients (right), shown with linear y-scale (top), square root y-scale (center) and logarithmic y-scale (bottom).

This algorithm separated the sources with a relative error of 0.67% for the least well separated source (error computed according to (19).) We also applied the BS Infomax algorithm (Bell and Sejnowski, 1995) implemented in Makeig (1999) to the spectrogram coefficients Y of the sensor signals. Separation errors were slightly larger, at 0.9%, but the computing time was improved (from 30 min for BFGS to 5 min for BS Infomax).

For comparison we tested the JADE (Cardoso, 1999a,b), FastICA (Hyvärinen, 1999, 1998) and BS Infomax algorithms on the raw sensor signals. Resulting relative errors (Figure 11) confirm the significant (by a factor of more than 10) superiority of the sparse decomposition approach.

The method described in this section, which combines a spectrogram transform with the BS Infomax algorithm, is included in the ICA/EEG toolbox (Makeig, 1999).

7 Future research

We should mention an alternative to the maximum a posteriori approach (12). Considering the mixing matrix A as a parameter, we can estimate it by maximizing the probability of the observed signal X

$$\max_A \left[P(X|A) = \int P(X|A,C)\,P(C)\,dC \right]$$

The integral over all possible coefficients C may be approximated, for example, by Monte-Carlo sampling or by a matching Gaussian, in the spirit of Lewicki and Sejnowski (1998); Lewicki and Olshausen (1999) or by variational methods (Jordan et al., 1999). It would be interesting to compare these possibility to the other methods presented in this paper.

Another important direction is towards the problem of simultaneous blind deconvolution and separation, as in Lambert (1996). In this case the matrices A and W will have linear

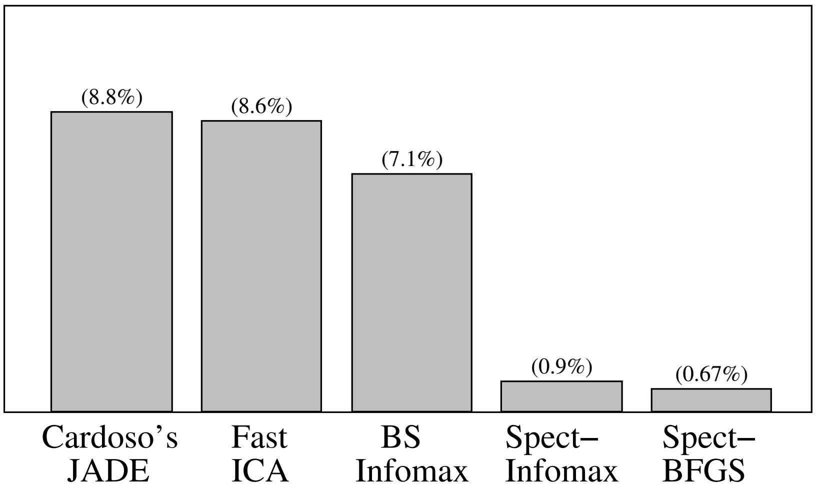

Figure 11: Percent relative error of separation of seven musical sources recovered by (1) JADE, (2) Fast ICA, (3) Bell-Sejnowski Infomax, (4) Infomax, applied to the spectrogram coefficients, (5) BFGS minimization of the objective (34) with the spectrogram coefficients.

filters as an elements, and multiplication by an element corresponds to convolution. Even in this matrix-of-filters context, most of the formulae in this paper remain valid.

8 Conclusions

We showed that the use of sparse decomposition in a proper signal dictionary provides high-quality blind source separation. The maximum *a posteriori* framework gives the most general approach, which includes the situation of more sources than sensors. Computationally more robust solutions can be found in the case of an equal number of sources and sensors. We can also extract the sources sequentially using quadratic programming with non-convex quadratic constraints. Finally, much faster solutions may be obtained by using non-overcomplete dictionaries. Our experiments with artificial signals and digitally mixed musical sounds demonstrate a high quality of source separation, compared to other known techniques.

Acknowledgments

This research was partially 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.

References

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

Attias, H. (1999). Independent factor analysis. *Neural Computation*, 11(4):803–851.

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

Belouchrani, A. and Cardoso, J.-F. (1995). Maximum likelihood source separation by the expectation-maximization technique: Deterministic and stochastic implementation. In *Proceedings of 1995 International Symposium on Non-Linear Theory and Applications*, pages 49–53, Las Vegas, NV. In press.

Ben-Tal, A. and Zibulevsky, M. (1997). Penalty/barrier multiplier methods for convex programming problems. *SIAM Journal on Optimization*, 7(2):347–366.

Buckheit, J., Chen, S. S., Donoho, D. L., Johnstone, I., and Scargle, J. (1995). About wavelab. Technical report, Department of Statistics, Stanford University. http://www-stat.stanford.edu/~donoho/Reports/.

Cardoso, J.-F. (1999a). High-order contrasts for independent component analysis. *Neural Computation*, 11(1):157–192.

Cardoso, J.-F. (1999b). JADE for real-valued data. http://sig.enst.fr/~cardoso/guidesepsou.html.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (1996). Atomic decomposition by basis pursuit. http://www-stat.stanford.edu/~donoho/Reports/.

Chen, S. S., Donoho, D. L., Saunders, M. A., Johnstone, I., and Scargle, J. (1995). About atomizer. Technical report, Department of Statistics, Stanford University. http://www-stat.stanford.edu/~donoho/Reports/.

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

Holmstrom, K. and Bjorkman, M. (1999). The TOMLAB NLPLIB. *Advanced Modeling and Optimization*, 1:70–86. http://www.ima.mdh.se/tom/.

Hyvärinen, A. (1998). The Fast-ICA MATLAB package. http://www.cis.hut.fi/~aapo/.

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

ICONIP'96 (1996). *International Conference on Neural Information Processing*, Hong Kong. Springer-Verlag.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. In Jordan, M. I., editor, *Learning in Graphical Models*, pages 105–161. Kluwer Academic Publishers.

Lambert, R. H. (1996). *Multichannel Blind Deconvolution: FIR Matrix Algebra and Separation of Multipath Mixtures*. PhD thesis, USC.

Lee, T. W., Lewicki, M. S., Girolami, M., and Sejnowski, T. J. (1998). Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Sig. Proc. Lett.* to appear.

Lewicki, M. S. and Olshausen, B. A. (1999). A probabilistic framework for the adaptation and comparison of image codes. *Journal of the Optical Society of America.* in press.

Lewicki, M. S. and Sejnowski, T. J. (1998). Learning overcomplete representations. *Neural Computation.* to appear.

Makeig, S. (1999). ICA/EEG toolbox. Computational Neurobiology Laboratory, the Salk Institute. http://www.cnl.salk.edu/˜tewon/ica_cnl.html.

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

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609.

Olshausen, B. A. and Field, D. J. (1997). Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Research*, 37:3311–3325.

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

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

Rowe, D. B. (1999). Bayesian blind source separation. *Submitted to IEEE Transactions on Signal Processing.*

Tang, A. C., Pearlmutter, B. A., and Zibulevsky, M. (1999). Blind separation of neuromagnetic responses. In *Computational Neuroscience.* Published in a special issue of *Neurocomputing.*

Tang, A. C., Pearlmutter, B. A., Zibulevsky, M., Hely, T. A., and Weisend, M. P. (2000a). An MEG study of response latency and variability in the human visual system during a visual-motor integration task. In *Advances in Neural Information Processing Systems 12*, pages 185–191. MIT Press.

Tang, A. C., Phung, D., Pearlmutter, B. A., and Christner, R. (2000b). Localization of independent components from magnetoencephalography. In *International Workshop on Independent Component Analysis and Blind Signal Separation*, Helsinki, Finland.

Zibulevsky, M., Pearlmutter, B. A., Bofill, P., and Kisilev, P. (2000). Blind source separation by sparse decomposition in a signal dictionary. In Roberts, S. J. and Everson, R. M., editors, *Independent Components Analysis: Princeiples and Practice.* Cambridge University Press. In press.