1

Blind Source Separation by Sparse Decomposition in a Signal Dictionary

M. Zibulevsky and B.A. Pearlmutter, University of New Mexico

P. Bofill, Universitat Politècnica de Catalunya

P. Kisilev, Technion—Israel Institute of Technology

1.1 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 $\xi(t)$

$$x(t) = As(t) + \xi(t) \tag{1.1}$$

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

$$s_i(t) = \sum_{k=1}^{K} C_{ik}\, \varphi_k(t) \tag{1.2}$$

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; Olshausen and Field, 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 $t = 1, 2, \ldots, T$ 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 (1.1) and (1.2) then take the following simple form

$$X = AS + \xi \tag{1.3}$$

$$S = C\Phi \tag{1.4}$$

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

$$X \approx AC\Phi$$

Our goal therefore can be formulated as follows:

Given sensor signal matrix X and 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† (Lewicki and Sejnowski, 1998; Lewicki and Olshausen, 1999; Olshausen 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.

Overview of the chapter

We start this chapter with some motivating examples, which demonstrate how sparsity helps to separate sources (Section 1.2). Then in Section 1.3 we present a *clustering* approach, which is one of the most efficient ways to estimate the mixing matrix when the sources are sparse.

Overcomplete dictionary. Section 1.4 gives the problem formulation in probabilistic framework in the most general case of an overcomplete dictionary, when there can be more sources than mixtures, and presents the *maximum a posteriori* approach to its solution.

In Section 1.5 we derive another objective function, which provides more robust computations when there are an equal number of sources

† Our dictionary Φ may be obtained in this way.

and mixtures. Section 1.6 presents sequential source extraction using quadratic programming with non-convex quadratic constraints.

Non-overcomplete dictionary. When the dictionary is non-overcomplete, computationally much faster solutions are possible. In Section 1.7 we demonstrate high-quality separation of synthetically mixed musical sounds with a square mixing matrix.

Even when the number of sources is larger than the number of mixtures, we can estimate the mixing matrix beforehand by clustering, and then reconstruct the sources by a *shortest path decomposition*, as it is shown in Section 1.8. Here we present examples of separation of up to six sound sources from two mixtures.

Exploiting multiscale representations In many cases, especially in wavelet-related decompositions, there are distinct groups of coefficients, in which sources have different sparsity properties. Section 1.9 shows, how selection of the best groups of coefficients significantly improves the separation quality.

1.2 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 1.4.1.

Example: 2 sources and 2 mixtures. Two synthetic sources are shown in Figure 1.1(a,b). The first source has two non-zero samples, and the second has three. The mixtures, shown in Figure 1.1(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 1.2(a), when μ is not optimal the second source interferes, and the total number of non-zero samples remains five. Only when the first

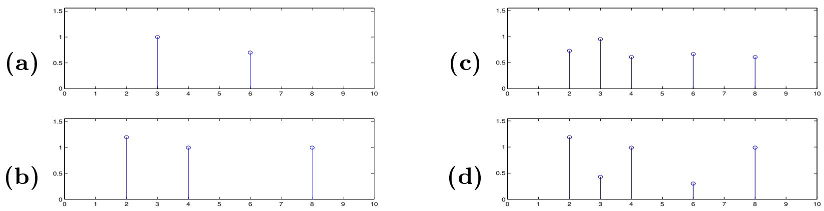


Fig. 1.1. Sources (a and b) are sparse. Mixtures (c and d) are less sparse.

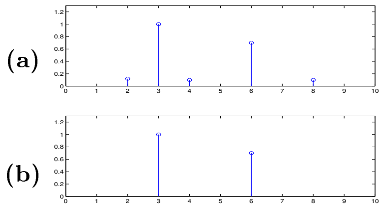


Fig. 1.2. (a) Imperfect separation. Since the second source is not completely removed, the total number of non-zero samples remains five. (b) Perfect separation. When the source is recovered perfectly, the number of non-zero samples drops to two and the objective function achieves its minimum.

source is recovered perfectly, as in Figure 1.2(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 1.3. 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 1.3 shows a scatter plot of the mixtures. The directions of the columns of mixing matrix are clearly visible. Indeed, if only one source, say $s_1(t)$, was present, the sensor signals would look like

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

and the points at the scatter plot of x_2 versus x_1 would belong to the straight line placed along the vector $[a_{11}a_{21}]^T$. The same thing hap-

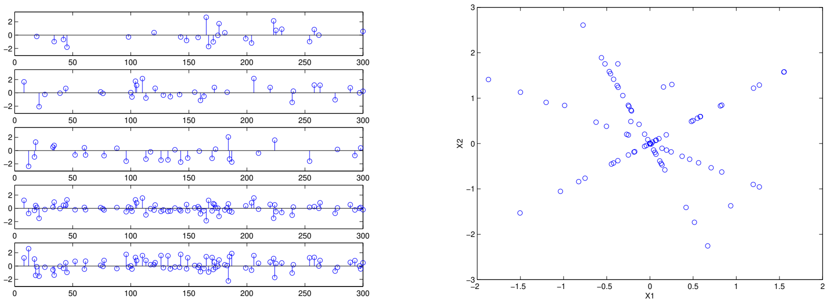


Fig. 1.3. Left: top three panels – sparse sources (sparsity is 10%), bottom two panels – mixtures. Right: scatter plot of two mixtures x_1 versus x_2. Three distinguished directions, which correspond to the columns of the mixing matrix $\mathbf{A}$, are visible.

pens when all the sources are present but the samples are sparse: at each particular index where a sample of one source is large, there is a high probability that the corresponding samples of other sources are small, and the point in the scatter plot still lies close to the mentioned straight line. This explains the appearance of dominant orientations at the scatter plot.

1.3 Clustering of Data Concentration Directions

The phenomena of data concentration along the directions of the columns of mixing matrix can be used in clustering approaches to source separation (Pajunen et al., 1996; Bofill and Zibulevsky, 2000b). This works efficiently even if the number of sources is greater than the number of sensors. In order to determine orientations of data concentration, we project the data points onto the surface of a unit sphere† by normalizing corresponding vectors, and then apply a standard clustering algorithm. Our clustering procedure can be summarized as follows:

(i) In order to project data points onto the surface of a unit sphere, normalize the sensor data vectors at every particular time index k: $\mathbf{x_k} = \mathbf{x_k}/||\mathbf{x_k}||$;

Before normalization, it is reasonable to remove data points with a very small norm, since these very likely are noisy.

† One can also use weights, depending on the distance of a data point from the origin, because more distant points are more reliable.

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

Without this operation each 'line' of data concentration would yield two clusters on opposite sides of the sphere.

(iii) Determine cluster centers using some clustering algorithm. Their coordinates will form the columns of the estimated mixing matrix $\tilde{\mathbf{A}}$.

In computational examples below in this chapter we use *C-means* clustering Bezdek, 1981 as implemented in the *Matlab Fuzzy Logic Toolbox* function FCM. We built also a modification of *C-means* algorithm, which allows its input points to be weighted. The optimal choice of the weights, as a function of the distance of a data point from the origin still requires further investigation. In Section 1.8 we use also *potential-function* based clustering Bofill and Zibulevsky, 2000b.

1.4 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 (1.4) are independent random variables with a probability density function (pdf) of an exponential type

$$p_i(C_{ik}) \propto \exp -\beta_i h(C_{ik}) \tag{1.5}$$

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

$$h(c) = |c|^{1/\gamma} \qquad\qquad \gamma \geq 1 \tag{1.6}$$

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

$$h_1(c) \quad = \quad |c| - \log(1 + |c|) \tag{1.7}$$

$$h_\lambda(c) \quad = \quad \lambda h_1(c/\lambda) \tag{1.8}$$

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 (1.3) is a spatially and temporally uncorrelated Gaussian process† with zero mean and variance σ^2.

1.4.1 Maximum a posteriori approach

We wish to maximize the posterior probability

$$\max_{A,C} P(A,C|X) \;\propto\; \max_{A,C} P(X|A,C)\, P(A)\, P(C) \tag{1.9}$$

where $P(X|A,C)$ is the conditional probability of observing X given A and C. Taking into account (1.3), (1.4), and the white Gaussian noise, we have

$$P(X|A,C) \propto \prod_{i,t} \exp -\frac{(X_{it} - (AC\Phi)_{it})^2}{2\sigma^2} \tag{1.10}$$

By the independence of the coefficients C_{jk} and (1.5), the prior pdf of C is

$$P(C) \propto \prod_{j,k} \exp(-\beta_j h(C_{jk})) \tag{1.11}$$

If the prior pdf $P(A)$ is uniform, it can be dropped† from (1.9). In this way we are left with the problem

$$\max_{A,C} P(X|A,C)\, P(C). \tag{1.12}$$

By substituting (1.10) and (1.11) into (1.12), taking the logarithm, and inverting the sign, we obtain the following optimization problem

$$\min_{A,C} \frac{1}{2\sigma^2} \|AC\Phi - X\|_F^2 + \sum_{j,k} \beta_j h(C_{jk}) \tag{1.13}$$

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; Olshausen and Field, 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 $C = 0$ and $A = \infty$. We can overcome this difficulty in various

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

† Otherwise, if $P(A)$ is some other known function, we should use (1.9) directly.

ways. The first approach is to force each row A_i of the mixing matrix A to be bounded in norm,

$$\|A_i\| \leq 1 \qquad\qquad i = 1, \ldots, N. \tag{1.14}$$

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

$$\|C_j\| \geq 1 \qquad\qquad j = 1, \ldots, M. \tag{1.15}$$

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 (1.5), 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,

$$\hat{\beta}_j = f_h^{-1}(K^{-1} \sum_k C_{jk}^2) \tag{1.16}$$

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

$$\hat{\beta}_j = \frac{2}{\sqrt{K^{-1} \sum_k C_{jk}^2}} \tag{1.17}$$

Substituting $h(\cdot)$ and $\hat{\beta}$ into (1.13), we obtain

$$\min_{A,C} \; \frac{1}{2\sigma^2} \|AC\Phi - X\|_F^2 + \sum_j \frac{2 \sum_k |C_{jk}|}{\sqrt{K^{-1} \sum_k C_{jk}^2}} \tag{1.18}$$

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

1.4.2 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 1.4. 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 1.5(a)), each composed

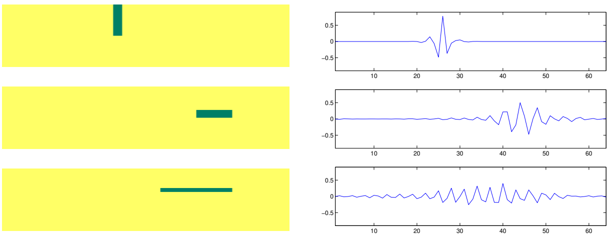
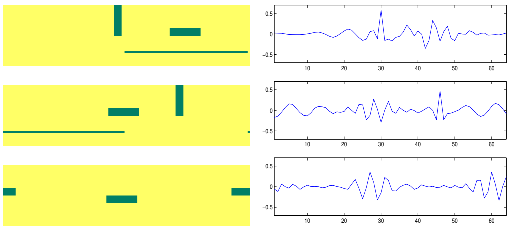
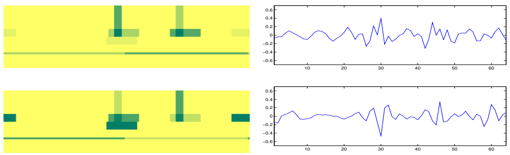


Fig. 1.4. Examples of atoms: time-frequency phase plane (left) and time plot (right.)

(a) Sources



(b) Mixtures



(c) Separated sources

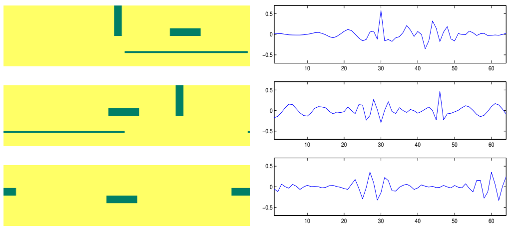


Fig. 1.5. (a) Sources, (b) mixtures, and (c) reconstructed sources, in both time-frequency phase plane (left) and time domain (right).

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 1.5(b)) were obtained as linear mixtures of the sources. In order to measure the accuracy of separation, we normalized the original sources with $\|S_j\|_2 = 1$, and the estimated sources with $\|\tilde{S}_j\|_2 = 1$. The error was computed as

$$\text{Error} = \frac{\|\tilde{S}_j - S_j\|_2}{\|S_j\|_2} \cdot 100\% \tag{1.19}$$

We tested two methods with this data. The first method used the objective function (1.13) and the constraints (1.15), while the second method used the objective function (1.18). 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 (1.7) and (1.8) 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 1.5(c). They are visually indistinguishable from the original sources in Figure 1.5(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 Bofill and Zibulevsky, 2000b.)

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.

1.5 Equal number of sources and sensors: more robust formulations

The main difficulty in a maximization problem like (1.13) 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 (1.4), we obtain a least squares term $\|C\Phi - WX\|_F^2$, so the separation objective may be written

$$\min_{W,C} \frac{1}{2}\|C\Phi - WX\|_F^2 + \mu \sum_{j,k} \beta_j h(C_{jk}) \tag{1.20}$$

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,

$$r_{\min}(W) \geq 1 \tag{1.21}$$

It can be shown that in the noiseless case, $\sigma \approx 0$, the problem (1.20)–(1.21) is equivalent to the maximum *a posteriori* formulation (1.13) with the constraint $\|A\|_2 \leq 1$. Another possibility for ensuring the non-singularity of W is to subtract $K \log|\det W|$ from the objective

$$\min_{W,C} -K \log|\det W| + \frac{1}{2}\|C\Phi - WX\|_F^2 + \mu \sum_{j,k} \beta_j h(C_{jk}) \tag{1.22}$$

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)

$$\min_{W} -K \log|\det W| + \sum_{j,k} \beta_j h((WX)_{jk}) \tag{1.23}$$

Experiment: equal numbers of sources and sensors. We created two sparse sources (Figure 1.6, top) with strong cross-correlation of 0.52. Separation by minimization of the objective function (1.22) gave an error of 0.23%. Robust convergence 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

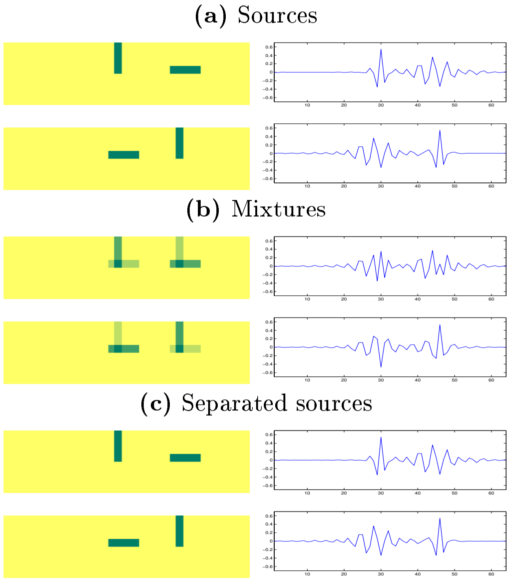


Fig. 1.6. (a) Sources, (b) mixtures, and (c) reconstructed sources, in both time-frequency phase plane (left) and time domain (right).

were used with default setting of all parameters. The resulting relative errors (Figure 1.7) 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.

1.6 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

$$\min_{w,c} \frac{1}{2} \|c^T \Phi - w^T X\|_2^2 + \mu \sum_k h(c_k), \tag{1.24}$$

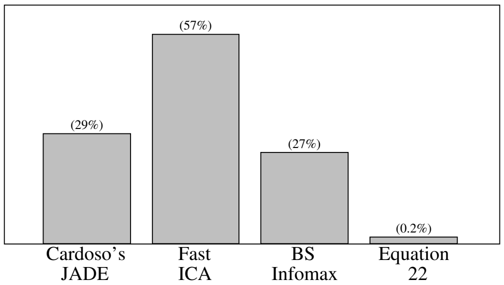


Fig. 1.7. Percent relative error of separation of the artificial sparse sources recovered by (1) JADE, (2) Fast ICA, (3) Bell-Sejnowski Infomax, (4) Equation 1.22.

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,

$$\|w\|_2^2 \geq 1 \ , \tag{1.25}$$

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 $w^1, \dots, w^{j-1}$

$$\|(I - P^{j-1})w^j\|_2^2 \geq 1 \ , \tag{1.26}$$

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

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 (1.24) and (1.26) 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.

1.7 Fast Solution in Non-overcomplete Dictionaries

In important applications (Tang et al., 1999; Tang et al., 2000), 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; Olshausen and Field, 1996).

Let us denote the dual basis

$$\Psi = \Phi^{-1} \tag{1.27}$$

and suppose that coefficients of decomposition of the sources

$$C = S\Psi \tag{1.28}$$

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

$$Y = X\Psi \tag{1.29}$$

Multiplying both sides of (1.3) by Ψ from the right and taking into account (1.28) and (1.29), we obtain

$$Y = AC + \zeta \ , \tag{1.30}$$

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 1.4.1–1.6 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 (1.13) and (1.18) are transformed into

$$\min_{A,C} \ \frac{1}{2\sigma^2}\|AC - Y\|_F^2 + \sum_{j,k} \beta_j h(C_{jk}) \tag{1.31}$$

and

$$\min_{A,C} \ \frac{1}{2\sigma^2}\|AC - Y\|_F^2 + \sum_{j} \frac{2\sum_k |C_{jk}|}{\sqrt{K^{-1}\sum_k C_{jk}^2}} \tag{1.32}$$

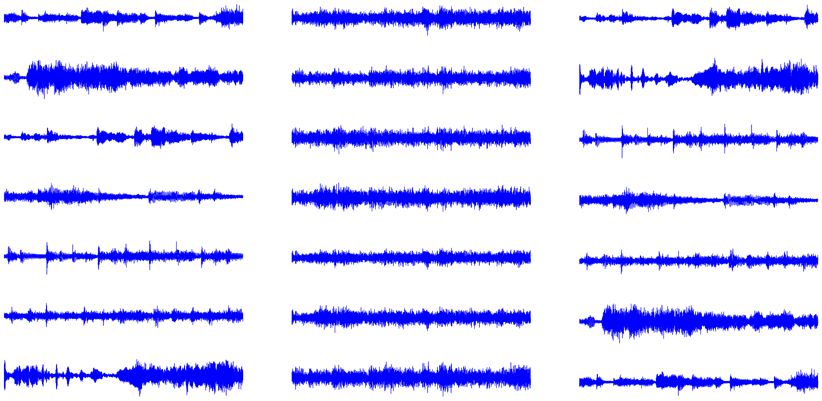


Fig. 1.8. Separation of musical recordings taken from commercial digital audio CDs (five second fragments).

The objective (1.22) becomes

$$\min_{W,C} \; -K \log|\det W| + \frac{1}{2}\|C - WY\|_F^2 + \mu \sum_{j,k} \beta_j h(C_{jk}) \qquad (1.33)$$

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)

$$\min_{W} \; -K \log|\det W| + \sum_{j,k} \beta_j h((WY)_{jk}) \qquad (1.34)$$

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 1.8).

The easiest way to perform sparse decomposition of such sources is to compute a *spectrogram*, the coefficients of a *Short Time Fourier Transform* (STFT). (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 1.9, right) is much higher then the sparsity of the original signal (Figure 1.9, left)

In this case Y (1.29) 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 (1.34) with $\beta_j = 1$ and $h_\lambda(\cdot)$ defined by (1.7) and (1.8) with the parameter $\lambda = 10^{-4}$. Unconstrained minimization was performed by a BFGS Quasi-Newton algorithm (MATLAB function FMINU.)

This algorithm separated the sources with a relative error of 0.67% for the least well separated source (error computed according to (1.19).) 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; Cardoso, 1999b), FastICA (Hyvärinen, 1999; Hyvärinen, 1998) and BS Infomax algorithms on the raw sensor signals. Resulting relative errors (Figure 1.10) 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).

1.8 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 $\mathbf{C} = \mathbf{W}\mathbf{Y}$, in the case of more sources than mixtures, we are faced with *two* interrelated problems: estimating the mixing matrix $\mathbf{A}$ *and* estimating the sources $\mathbf{C}$. Trying to solve both of them at the same time as in equation (1.31) is a difficult multivariate optimization problem.

Another approach consists in estimating the mixing matrix $\mathbf{A}$ beforehand. We can do this by clustering (as in Section 1.3), using sparsity of sensor coefficients $\mathbf{Y}$. In experiments of this section we use sparsity

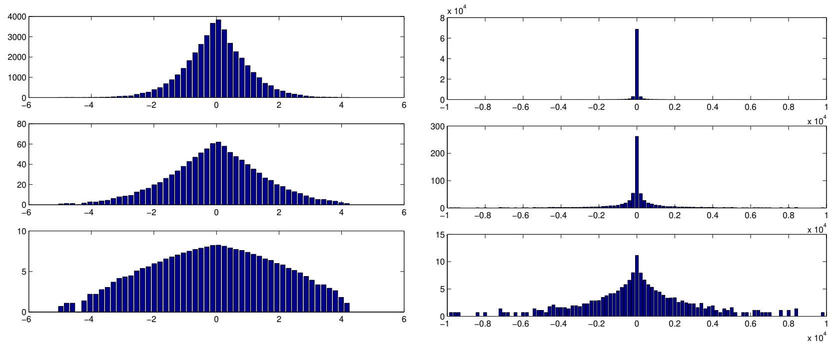


Fig. 1.9. Histogram of sound source values (left) and spectrogram coefficients (right), shown with linear y-scale (top), square root y-scale (center) and logarithmic y-scale (bottom).

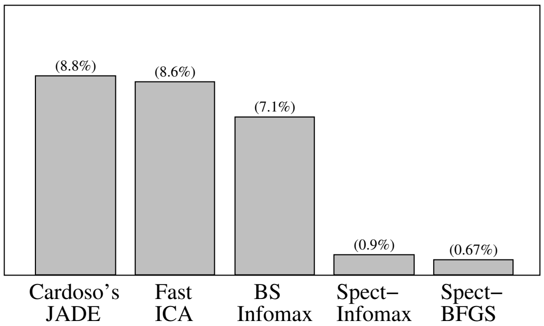


Fig. 1.10. Percent relative error of separation of seven musical sources recovered by (1) JADE, (2) Fast ICA, (3) Bell-Sejnowski Infomax, (4) Infomax, applied to the spectrogram coefficients, (5) BFGS minimization of the objective (1.34) with the spectrogram coefficients.

of *Short Time Fourier Transform* (STFT). The benefits of such an approach are clear in Figure 1.11. Six flute signals playing different notes (see the *Six Flutes* example in Section 1.8.2) were synthetically mixed into two mixtures along equally spaced directions. Figure 1.11a 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 1.11b (i.e., x_2^w against x_1^w for every w). The difference is extraordinary. Now almost

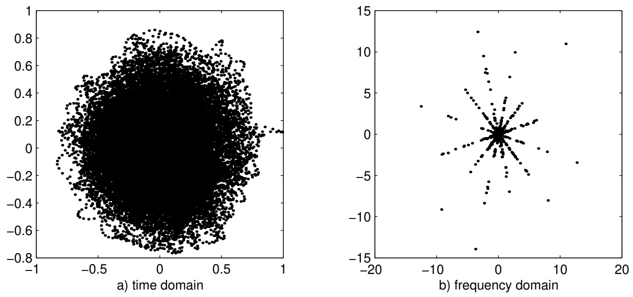


Fig. 1.11. Scatter plot $\mathbf{X}_{2\cdot}$ vs $\mathbf{X}_{1\cdot}$ of six flute notes mixed into two mixtures along equally spaced directions in the time (left) and frequency (right) domains.

all the data points are neatly clustered along the six directions of the columns of the mixing matrix, thus providing very good separability.

If we assume that the matrix $\mathbf{A}$ is found, the problem (1.31) can be decomposed into to K independent small problems for each data point c^k (here we use $h(\cdot) = |\cdot|$)

$$\min_{\mathbf{c}^k} \frac{1}{2\sigma^2}||\mathbf{A}\mathbf{c}^k - \mathbf{y}^k||^2 + \sum_j |c_j^k|, \quad \text{for } k = 1, \ldots, K. \tag{1.35}$$

Or, in the absence of noise

$$\min_{\mathbf{c}^k} \sum_j |c_j^k| \quad \text{subject to} \quad \mathbf{A}\mathbf{c}^k = \mathbf{y}^k, \quad \text{for } k = 1, \ldots, K, \tag{1.36}$$

which can be formulated as a linear programming problem Chen et al., 1996.

1.8.1 A Shortest Path Decomposition of the Sources

We use a simple geometrical approach to the optimization problem (1.36). When the columns $\mathbf{a}^j$ are normalized, the optimal representation of the data point $\mathbf{y}^k = \sum_j \mathbf{a}^j c_j^k$ that minimizes $\sum_j |c_j^k|$, will include at most N of the $\mathbf{a}^j$'s, corresponding to the vertices of the *minimal simplex* enclosing the direction of vector $\mathbf{y}^k$ (this leads to the problem of triangulation on sphere.) 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 included into the path.

In particular, for the two-sensor case, the shortest path is obtained

Six Flutes (FFT)	50.5	52.5	49.4	43.4	49.1	51.8
Six Flutes (time domain)	-1.9	-2.0	-2.2	-2.4	-2.3	-2.4
Four Voices (STFT)	21.7	19.4	15.7	16.6		
Five Songs (STFT)	15.6	15.5	15.0	15.1	15.2	
Six Flute Melodies (STFT)	20.4	19.4	14.2	16.1	24.7	29.1

Table 1.1. *S/N reconstruction indices (dB) for the different experiments (see text).*

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 of the data point $\theta_k = \tan^{-1}(y_2^k/y_1^k)$.

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

$$\begin{aligned} \mathbf{c}_r^k &= \mathbf{W}_r\mathbf{y}^k, \\ c_j^k &= 0, \quad \text{for} \quad j \neq b, a. \end{aligned} \tag{1.37}$$

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

1.8.2 Experiments with Estimating the Mixing Matrix and the Sources Separately

The approach was first tested using the *Six Flutes* 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. 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 mixed into two mixtures along equally spaced directions. 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. We used *potential function* based clustering Bofill and Zibulevsky, 2000b. Results are shown in the first row of Table 1.1.

For the sake of comparison, the next experiment was conducted on the same data set using the mixtures in the time domain instead of in the frequency domain. The centers of obtained clusters 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 1.1, the second row.

The flute notes in the *Six Flutes* 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 STFT. As before, the sources were first normalized to the same energy level and mixed in the time domain. STFT of the resulting mixtures was produced with a Hanning window of length L, and a "hop" distance d was used between the starting point of successive frames (yielding an $L-d$ overlap). For each mixture, the input to the separation system was then a single long vector containing the concatenation of the coefficients of real and imaginary parts of the positive spectra among all the frames in that mixture. After the separation 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 *Four Voices* 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. STFT was done with $L = 2048$ and $d = 614$ samples. A *Five Songs* 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 *Six Flute Melodies* data set 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. Results of the separation are shown in Table 1.1. Although good enough in themselves, the reconstruction indices of the dynamic signals were significantly poorer than those of the *Six Flutes*, in part due to the intrinsic difficulties of the short-term analysis and resynthe-

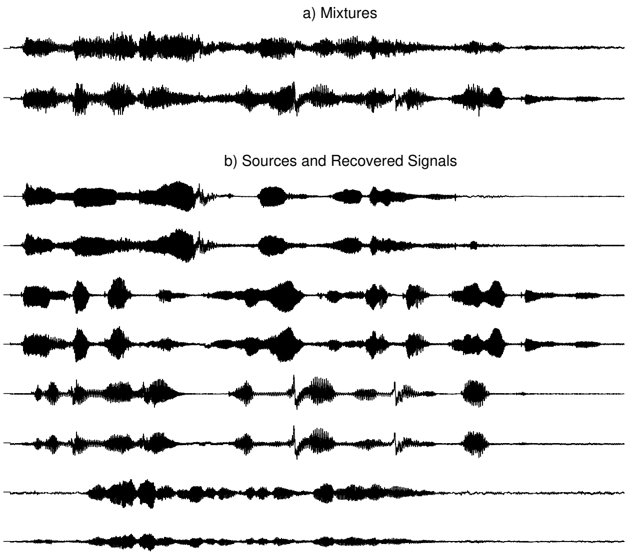


Fig. 1.12. FourVoices experiment. (a) Mixtures, (b) sources and recovered signals, pairwise. Taken from Bofill and Zibulevsky, 2000a.

sis. 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 1.12 for the *Four Voices* 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. Sound examples for the above experiments are available on-line at http://www.ac.upc.es/homes/pau/.

1.9 Source Separation Using Sparsity of Multiscale Representation

In many cases, especially in wavelet-related decompositions, there are distinct groups of coefficients, in which sources have different sparsity properties. The idea is to select those groups of features (coefficients) which are best suited for separation, with respect to the following criteria: (1) sparsity of coefficients (2) separability of sources' features. After the best groups are selected, one uses only these in the separation process, which can be accomplished by standard ICA algorithms or by clustering. We present experiments with simulated signals, musical sounds and images which demonstrate improvement of separation quality.

1.9.1 Example: sparsity of random blocks in the Haar basis

Typical block functions are shown in Figure 1.13. They are piecewise constant, with random amplitude and duration of each constant piece. Let us take a close look at the Haar wavelet coefficients at different resolutions. Wavelet basis functions at the finest resolution are obtained by translation of the Haar mother wavelet:

$$\varphi_j(t) = \begin{cases} -1 & \text{if } t = 0 \\ 1 & \text{if } t = 1 \\ 0 & \text{otherwise .} \end{cases}$$

Taking a scalar product of a function $s(t)$ with the wavelet $\varphi_j(t-\tau)$, we produce a finite differentiation of the function $s(t)$ at the point $t = \tau$. This means that the number of non-zero coefficients at the finest resolution for a block function will correspond roughly to the number of jumps it has. Proceeding to the next, coarser resolution level

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

the number of non-zero coefficients still corresponds to the number of jumps, but the total number of coefficients at this level is halved , and so is the sparsity. If we proceed further in this direction, we will achieve levels of resolution, where typical width 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 will fade-out.

To demonstrate how this influences accuracy of a blind source separation, we randomly generated two block-signal sources (Fig 1.13, left), and mixed them by the matrix

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

The resulting mixtures, $x_1(t)$ and $x_2(t)$ are shown in Figure 1.13, center. Figure 1.14, first column, shows the scatter plot of $x_1(t)$ versus $x_2(t)$,

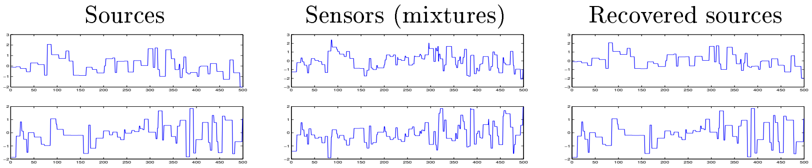
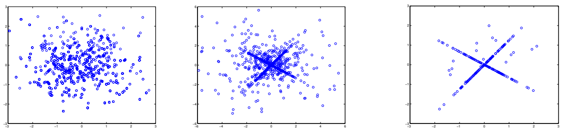


Fig. 1.13. Time plots of block signals



	Raw signals	All wavelet coefficients	Hi resolution wavelet coefficients
BS-Infomax	13.9%	4.2%	0.69%
C-means clustering	13.3%	2.4%	0.41%

Fig. 1.14. Separation of block signals: scatter plots of sensor signals and mean-squared separation errors (%)

where there are no visible distinct features. In contrast, the scatter plot of the wavelet coefficients at the highest resolution (Figure 1.14, third column) shows two distinct orientations, which correspond to the columns of the mixing matrix.

Results of separation of the block sources are presented in Figure 1.14. The largest error (13%) was obtained on the raw data, and the smallest (below 0.7%) – on the wavelet coefficients at the highest resolution, which have the best sparsity. Use of all wavelet coefficients leads to intermediate sparsity and performance.

1.9.2 Adaptive selection of sparse subsets of coefficients in wavelet packets tree

Multiresolution analysis

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 basis functions, for example, wavelet packets Coifman et al., 1992,

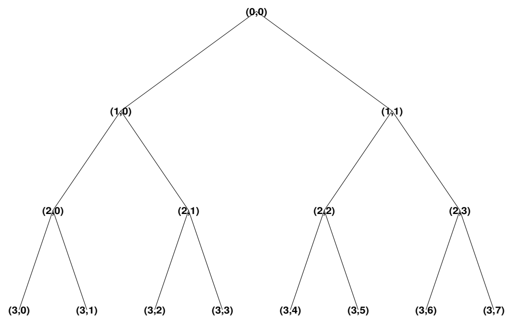


Fig. 1.15. Wavelet packets tree

or multiwavelets Weitzer et al., 1997, might be more appropriate. The wavelet packets library consists of the triple-indexed family of functions:

$$\varphi_{jnk}(t) = 2^{j/2}\varphi_n(2^j t - k),\ j,k \in Z,\, n \in N. \tag{1.38}$$

As in the case of the wavelet transform, j,k are the scale and shift parameters, respectively, and n is the frequency parameter, related to the number of oscillations of a particular generating function $\varphi_n(t)$. The set of functions $\varphi_{jn}(t)$ forms a (j,n) wavelet packet. This set of functions can be split into two parts at a coarser scale: $\varphi_{j-1,2n}(t)$ and $\varphi_{j-1,2n+1}(t)$. It follows that these two form an orthonormal basis of the subspace which spans $\{\varphi_{jn}(t)\}$. Thus, we arrive at a family of wavelet packet functions on a binary tree (Figure 1.15). The nodes of this tree are numbered by two indices: the depth of the level $j = 0,1,..,J$, and the number of nodes $n = 0,1,2,3,...,2^j{-}1$ at the specified level. Using wavelet packets allows one to analyze given signals not only with a scale-oriented decomposition but also on frequency sub-bands. Naturally, the library contains the wavelet basis.

The decomposition coefficients $c_{jnk} = \langle s, \varphi_{jnk}\rangle$ also split into (j,n) sets corresponding to the nodes of the tree, and there is a fast way to compute them using banks of *conjugate mirror filters*, as is implemented in the fast wavelet transform.

Choice of the best nodes in the tree

When signals have a complex nature, it is difficult to decide in advance which nodes contain the sparsest sets of coefficients. That is why we use the following simple adaptive approach.

First, for every node of the tree, we apply a clustering algorithm (see Section 1.3), 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). Second, we choose a few best nodes with the minimal distortion, combine their coefficients into one data set, and apply a separation algorithm (clustering or Infomax) to these data.

More sophisticated techniques dealing with adaptive choice of best nodes, as well as their number can be found in Kisilev et al., .

1.9.3 Experiments with adaptive selection of sparse subsets of coefficients

We evaluated the quality of the proposed wavelet-packet based separation method on several types of signals. The first type is the random blocks signal (see above). The second type of signal is a frequency modulated (FM) sinusoidal signal. In the first case, the carrier is modulated by a sinusoidal function. In the second case, it is modulated by choosing a random frequency and a corresponding random duration; we call this type of signal Block-FM (BFM). The third type of signal is a musical recording of flute sounds. Finally, we apply our algorithm to portrait images.

In order to compare the accuracy of our method to other methods, we form the following features sets: (1) the set of signals, (2) short time Fourier transform (STFT) coefficients, (3) Wavelet transform coefficients, and (4) Wavelet packets coefficients at the "best" nodes. In the last case, mixtures of sources were decomposed with the *Matlab wavelet packet toolbox* using various families of mother wavelets with different numbers of vanishing moments (smoothness parameter). A typical example of scatter plots of the wavelet packet coefficients at different nodes of the wavelet packet tree is shown in Figure 1.16. The upper left scatter plot, labeled "C", corresponds to the set of coefficients at all nodes. The reminder are the scatter plots of sets of coefficients indexed in a wavelet packet tree above. Generally speaking, the more distinct the directions appearing on these plots, the more precise the estimation of the mixing matrix, and, therefore, the better the separation.

We applied the fuzzy C-means clustering algorithm with some modifications (see Kisilev et al., for details) to each feature set. Table 1.13 summarizes results of our experiments. We compared the quality of separation of random block and BFM signals by performing 100 Monte-

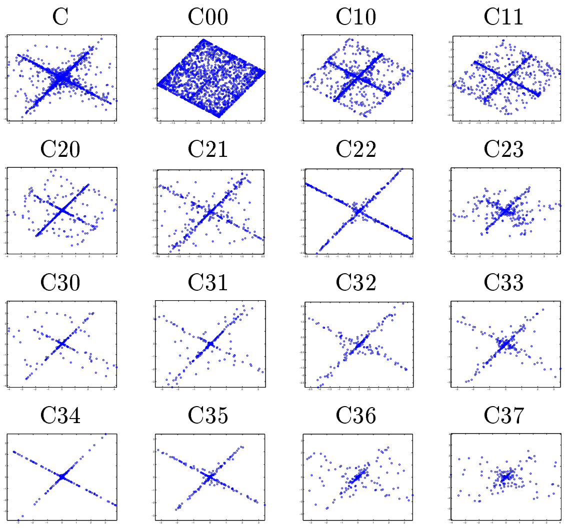


Fig. 1.16. Scatter plots of the WP coefficients of the FM mixtures

Signal	raw data	STFT	WT, db8	WT, haar	WP, db8	WP, haar
Blocks	31.89	16.31	4.18	1.94	2.70	**0.43**
BFM sine	49.81	8.17	8.16	15.30	**4.48**	6.65
FM sine	50.57	5.66	10.16	24.71	**4.13**	5.33
Flutes	12.18	5.36	5.96	9.23	**3.93**	8.05
Images	raw data	DCT	WT, sym8	WT, haar	WP, sym8	WP, haar
Portraits	22.11	19.11	10.79	10.57	**6.04**	8.29

Table 1.2. *Experimental results: normalized mean square separation error (%) for signals and images using raw data and decomposition coefficients in different domains. In the case of wavelet packets (WP) we used the best selected nodes.*

Carlo simulations and calculating the normalized mean-squared errors (NMSE) for the above features sets. In the case of deterministic signals, we calculated a normalized squared error (SE). In the case of image separation, we used the 2D Discrete Cosine Transform (DCT) instead of the STFT, and the *Symmlet-8* mother wavelet when using 2D wavelet transform and wavelet packets.

From Table 1.13 it is clear that the adaptive best nodes method outperforms all other feature sets for each type of signal. Also, as mentioned above, the clustering approach provides a better separation than InfoMax. It is clear that using the Haar wavelet function for the wavelet packets representation of the random block signals provides better separation than using some smooth wavelet, *e.g.* Db8. The reason is that these signals have a sparser representation with the Haar wavelet. In contrast, the Flute's signals are better represented with smooth wavelets, and, therefore, these provide yield 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 a number decomposition trees simultaneously.

More results and comparisons can be found in Kisilev et al., .

1.10 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 overcomplete dictionary and 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.

Much faster solutions may be obtained by using non-overcomplete dictionaries. Even when the number of sources is larger than the number of mixtures, we can estimate the mixing matrix beforehand by clustering, and then reconstruct the sources by a *shortest path decomposition.*

In many cases, especially in wavelet-related decompositions, selection of few best groups of coefficients with the highest sparsity brings additional improvement of the separation quality.

Our experiments with artificial signals and digitally mixed musical sounds demonstrate a high quality of source separation, compared to other known techniques.

1.11 Acknowledgements

We thank Linda Antas at the University of Washington for the flute performances, 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.

Bibliography

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.

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.

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

Bofill, P. and Zibulevsky, M. (2000a). Blind separation of more sources than mixtures using the sparsity of the short-time fourier transform. In *International Workshop on Independent Component Analysis and Blind Signal Separation*, Helsinki, Finland. In press.

Bofill, P. and Zibulevsky, M. (2000b). Sparse underdetermined ICA: Estimating the mixing matrix and the sources separately. Technical Report UPC-DAC-2000-7, Universitat Politecnica de Catalunya. http://www.ac.upc.es/homes/pau/sounds.html.

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:80/~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., Meyer, Y., and Wickerhauser, M. V. (1992). Wavelet analysis and signal processing. *in Wavelets and their applications, Jones and Barlett. Ruskai B. et al. eds., Boston.*

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.

Kisilev, P., Zibulevsky, M., Zeevi, Y. Y., and Pearlmutter, B. A. Multiresolution framework for sparse blind source separation. Technical report. in preparation.

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.

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

Tang, A. C., Pearlmutter, B. A., Zibulevsky, M., Hely, T. A., and Weisend, M. P. (2000). 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. In press.

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