Blind Source Separation by Sparse Decomposition in a Signal Dictionary

Michael Zibulevsky
Department of Computer Science, University of New Mexico, Albuquerque, NM 87131, U.S.A.

Barak A. Pearlmutter
Department of Computer Science and Department of Neurosciences, University of New Mexico, Albuquerque, NM 87131, U.S.A.

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, and other areas. We suggest a two-stage separation process: 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, followed by 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 nonovercomplete dictionary and an equal numbers of sources and mixtures. Experiments with artificial signals and 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 $\xi(t)$,

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

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 (e.g., wavelet packets, stationary wavelets; see Chen, Donoho, & Saunders, 1996; Mallat, 1998) and learned dictionaries (Lewicki & Sejnowski, in press; Lewicki & Olshausen, 1998; Olshausen & Field, 1996, 1997). Sparsity means that only a small number of the coefficients C_{ik} differ significantly from zero.

We suggest a two-stage separation process: a priori selection of a possibly overcomplete signal dictionary in which the sources are assumed to be sparsely representable and then 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 ith 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 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 Chen et al., 1996). Another problem is to adapt the dictionary to the given class of signals[1] (Lewicki & Sejnowski, 1998; Lewicki & Olshausen, 1998; Olshausen & Field, 1997). This problem is shown to be equivalent to the problem of blind source separation when the sources are sparse in time (Lee, Lewicki, Girolami, & Sejnowski, 1999; Lewicki & Sejnowski, in press). Our problem is different, but we will use and generalize some techniques presented in these works.

Independent factor analysis (Attias, 1999) and 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 the extreme case, when the decomposition coefficients are very sparse, the separation becomes practically ideal (see section 3.2 and the six flutes example in Zibulevsky, Pearlmutter, Bofill, & Kisilev, in press). Nevertheless, detailed comparison of the methods on real-world signals remains open for future research.

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

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 is an equal number of sources and mixtures. Section 5 presents sequential source extraction using quadratic programming with nonconvex quadratic constraints. Finally, in section 6 we derive a faster method for nonovercomplete dictionaries and demonstrate high-quality separation of synthetically mixed musical sounds.

2 Separation of Sparse Signals

In this section we present two examples that 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.

2.1 Example: Two Sources and Two Mixtures. Two synthetic sources are shown in Figures 1a and 1b. The first source has two nonzero samples, and the second has three. The mixtures, shown in Figures 1c and 1d, are less sparse: they have five nonzero samples each. One can use this observation to recover the sources. For example, we can express one of the sources as

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

and choose μ so as to minimize the number of nonzero samples $\|\widetilde{s}_i\|_0$, that is, the l_0 norm of s_i.

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

Note that the function $\|\widetilde{s}_i\|_0$ 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 nonzero. Fortunately in many cases, the l_1 norm $\|\widetilde{s}_i\|_1$ is a good substitute for this objective function. In this example, it too yields perfect separation.

2.2 Example: Three Sources and Two Mixtures. The signals are presented in Figure 3. These sources have about 10% nonzero samples. The nonzero 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.

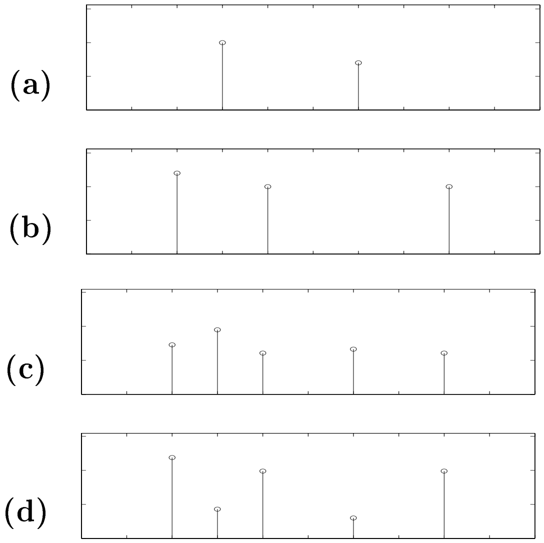

Figure 1: Coefficients of signals, with coefficient identity on the *x*-axis (10 coefficients, arbitrarily ordered) and magnitude on the *y*-axis (arbitrarily scaled). Sources (a and b) are sparse. Mixtures (c and d) are less sparse.

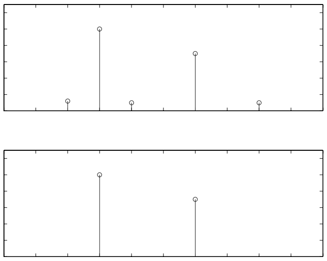

Figure 2: Coefficients of signals, with coefficient identity on the *x*-axis (10 coefficients, arbitrarily ordered) and magnitude on the *y*-axis (arbitrarily scaled). (a) Imperfect separation. Since the second source is not completely removed, the total number of nonzero samples remains five. (b) Perfect separation. When the source is recovered perfectly, the number of nonzero samples drops to two, and the objective function achieves its minimum.

This phenomenon can be used in clustering approaches to source separation (Pajunen, Hyvrinen, & Karhunen, 1996; Zibulevsky et al., in press). In this work we will explore a maximum a posteriori approach.

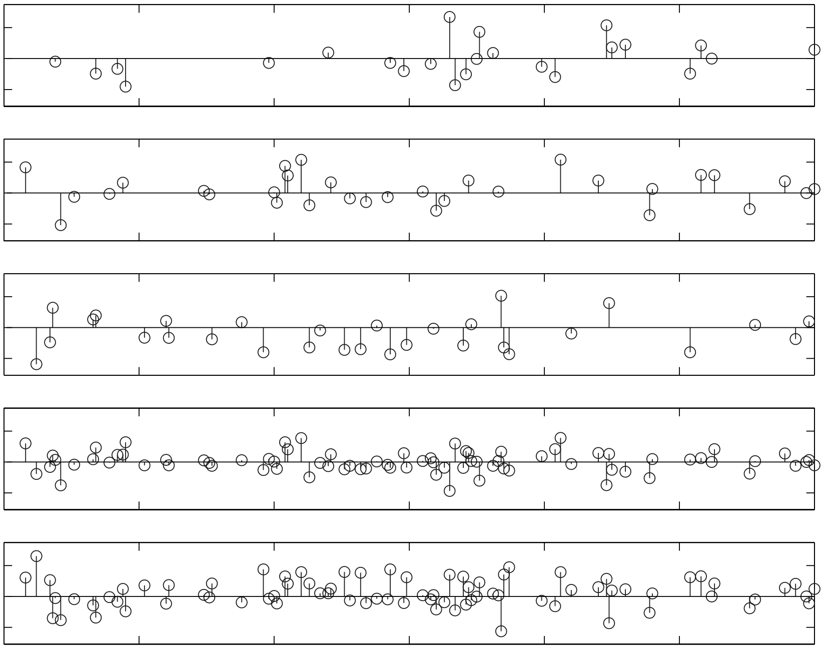

Figure 3: Coefficients of signals, with coefficient identity on the x-axis (300 coefficients, arbitrarily ordered) and magnitude on the y-axis (arbitrarily scaled). (Top three panels) Sparse sources (sparsity is 10%). (Bottom two panels) Mixtures.

3 Probabilistic Framework

In order to derive a maximum a posteriori solution, we consider the blind source separation problem in a probabilistic framework (Belouchrani & Cardoso, 1995; Perlmutter & Parra, 1996). Suppose that the coefficients C_{ik} in a source decomposition (see equation 1.4) are independent random variables with a probability density function (pdf) of an exponential type,

$$p_i(C_{ik}) \propto \exp -\beta_i h(C_{ik}). \tag{3.1}$$

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

$$h(c) = |c|^{1/\gamma} \qquad\qquad \gamma \geq 1 \tag{3.2}$$

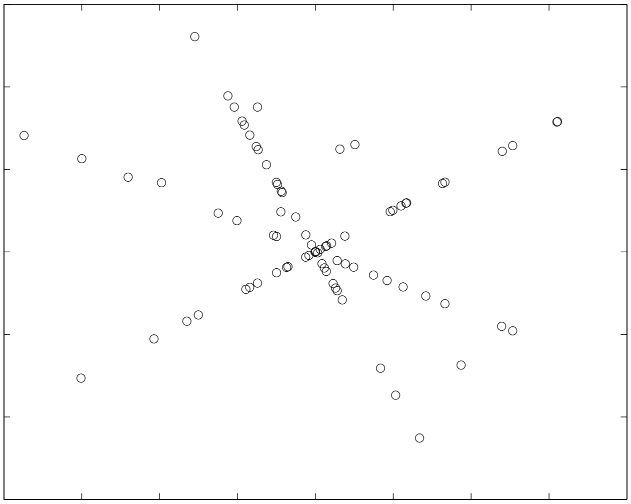

Figure 4: Scatter plot of two sensors. Three distinguished directions, which correspond to the columns of the mixing matrix A, are visible.

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

$$h_1(c) = |c| - \log(1 + |c|) \tag{3.3}$$

$$h_\lambda(c) = \lambda h_1(c/\lambda), \tag{3.4}$$

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 equation 1.3 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,

$$\max_{A,C} P(A, C|X) \propto \max_{A,C} P(X|A, C)\, P(A)\, P(C), \tag{3.5}$$

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

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

we have

$$P(X|A,C) \propto \prod_{i,t} \exp -\frac{(X_{it} - (AC\Phi)_{it})^2}{2\sigma^2}. \tag{3.6}$$

By the independence of the coefficients C_{jk} and equation 3.1, the prior pdf of C is

$$P(C) \propto \prod_{j,k} \exp(-\beta_j h(C_{jk})). \tag{3.7}$$

If the prior pdf $P(A)$ is uniform, it can be dropped[3] from equation 3.5. In this way we are left with the problem

$$\max_{A,C} P(X|A,C)\, P(C). \tag{3.8}$$

By substituting 3.6 and 3.7 into 3.8, taking the logarithm, and inverting the sign, we obtain the following optimization problem,

$$\min_{A,C} \frac{1}{2\sigma^2} \|AC\Phi - X\|_F^2 + \sum_{j,k} \beta_j h(C_{jk}), \tag{3.9}$$

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 $C = 0$ 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,

$$\|A_i\| \le 1 \qquad\qquad i = 1, \ldots, N. \tag{3.10}$$

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

$$\|C_j\| \ge 1 \qquad\qquad j = 1, \ldots, M. \tag{3.11}$$

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 3.1, express the variance of C_{jk} as

[3] Otherwise, if $P(A)$ is some other known function, we should use equation 3.5 directly.

a function $f_h(\beta)$. An estimate of β can be obtained by applying the corresponding inverse function to the sample variance,

$$\hat{\beta}_j = f_h^{-1}\left(K^{-1}\sum_k C_{jk}^2\right). \tag{3.12}$$

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

$$\hat{\beta}_j = \frac{2}{\sqrt{K^{-1}\sum_k C_{jk}^2}}. \tag{3.13}$$

Substituting $h(\cdot)$ and $\hat{\beta}$ into equation 3.9, we obtain

$$\min_{A,C} \frac{1}{2\sigma^2}\|AC\Phi - X\|_F^2 + \sum_j \frac{2\sum_k |C_{jk}|}{\sqrt{K^{-1}\sum_k C_{jk}^2}}. \tag{3.14}$$

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 that 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; it is overcomplete by a factor of seven. Examples of atoms and their images in the time-frequency phase plane (Coifman & Wickerhauser, 1992; Mallat, 1998) are shown in Figure 5. We used the ATOMIZER (Chen, Donoho, Saunders, Johnstone, & Scargle, 1995) and WAVELAB (Buckheit, Chen, Donoho, Johnstone, & Scargle, 1995) MATLAB packages for fast multiplication by Φ and Φ^T.

We created three very sparse sources (see Figure 6a), 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 (see Figure 6b) 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 $\|\widetilde{S}_j\|_2 = 1$. The error was computed as

$$\text{Error} = \frac{\|\widetilde{S}_j - S_j\|_2}{\|S_j\|_2} \cdot 100\%. \tag{3.15}$$

We tested two methods with these data. The first method used the objective function (see equation 3.9) and the constraints (see equation 3.11),

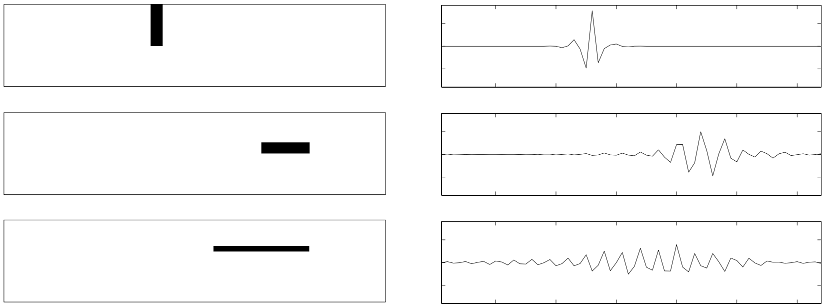

Figure 5: Examples of atoms. (Left) Time-frequency phase plane. (Right) Time plot.

and the second method used the objective function (see equation 3.14). We used PBM (Ben-Tal & Zibulevsky, 1997) for the constrained optimization. The unconstrained optimization was done using the method of conjugate gradients, with the TOMLAB package (Holmstrom & Bjorkman, 1999). The same tool was used by PBM for its internal unconstrained optimization.

We used $h_\lambda(\cdot)$ defined by equations 3.3 and 3.4, 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 6c. They are visually indistinguishable from the original sources in Figure 6a.

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% to 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., in press).

Second, the method of conjugate gradients requires a few thousand iterations to converge, which takes about 5 minutes 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 (2000) block coordinate descent method may be appropriate.) Below we present a few other approaches that help to stabilize and accelerate the optimization.

4 Equal Number of Sources and Sensors: More Robust Formulations

The main difficulty in a maximization problem like equation 3.9 is the bilinear term $AC\Phi$, which destroys the convexity of the objective function and

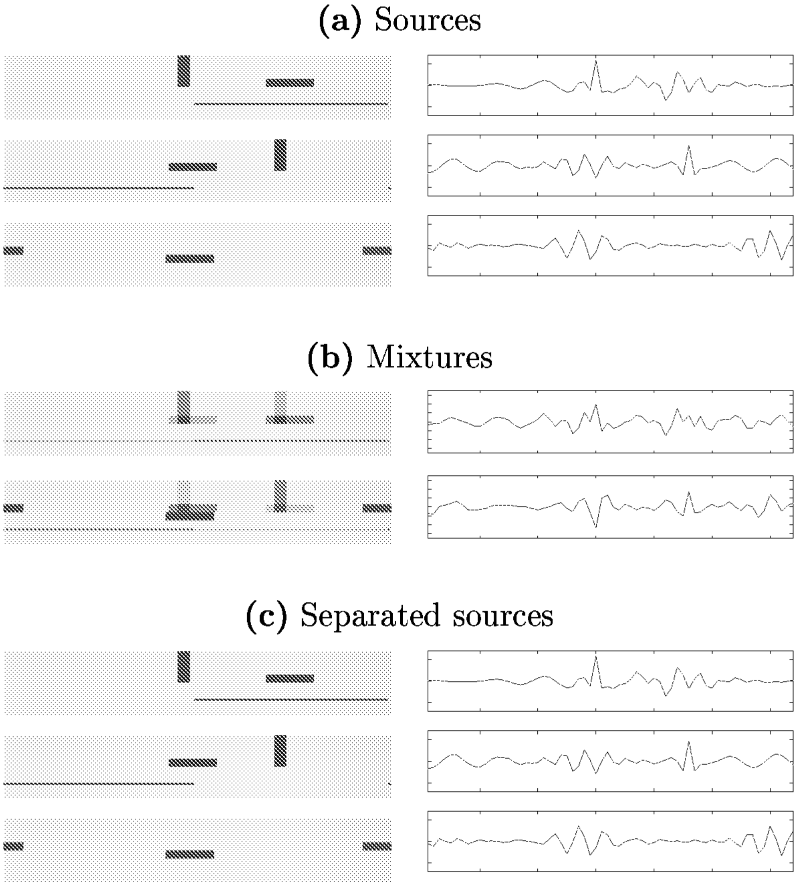

Figure 6: Sources, mixtures, and reconstructed sources in both time-frequency phase plane (left) and time domain (right).

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 equation 1.4, we obtain a least-squares term $\|C\Phi - WX\|_F^2$, so the separation objective may be written as

$$\min_{W,C} \frac{1}{2}\|C\Phi - WX\|_F^2 + \mu \sum_{j,k} \beta_j h(C_{jk}). \tag{4.1}$$

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

$$r_{\min}(W) \geq 1. \tag{4.2}$$

It can be shown that in the noiseless case, $\sigma \approx 0$, the problem 4.1–4.2 is equivalent to the maximum a posteriori formulation, equation 3.9, with the constraint $\|A\|_2 \leq 1$. Another possibility for ensuring the nonsingularity of W is to subtract $K \log |\det W|$ from the objective

$$\min_{W,C} -K \log |\det W| + \frac{1}{2}\|C\Phi - WX\|_F^2 + \mu \sum_{j,k} \beta_j h(C_{jk}) \tag{4.3}$$

which (Bell & Sejnowski, 1995; Pearlmutter & Para, 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 & Sejnowski, 1995):

$$\min_{W} -K \log |\det W| + \sum_{j,k} \beta_j h((WX)_{jk}). \tag{4.4}$$

4.1 Experiment: Equal Numbers of Sources and Sensors. We created two sparse sources (see Figure 7, top) with strong cross-correlation of 0.52. Separation by minimization of the objective function, equation 4.3, 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 & Sejnowski, 1995; Amari, Cichocki, & Yang, 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 (see Figure 8) confirm the significant superiority of the sparse decomposition approach.

This still takes a few thousand conjugate gradient steps to converge (about 5 minutes 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.

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

$$\min_{w,c} \frac{1}{2}\|c^T\Phi - w^T X\|_2^2 + \mu \sum_k h(c_k), \tag{5.1}$$

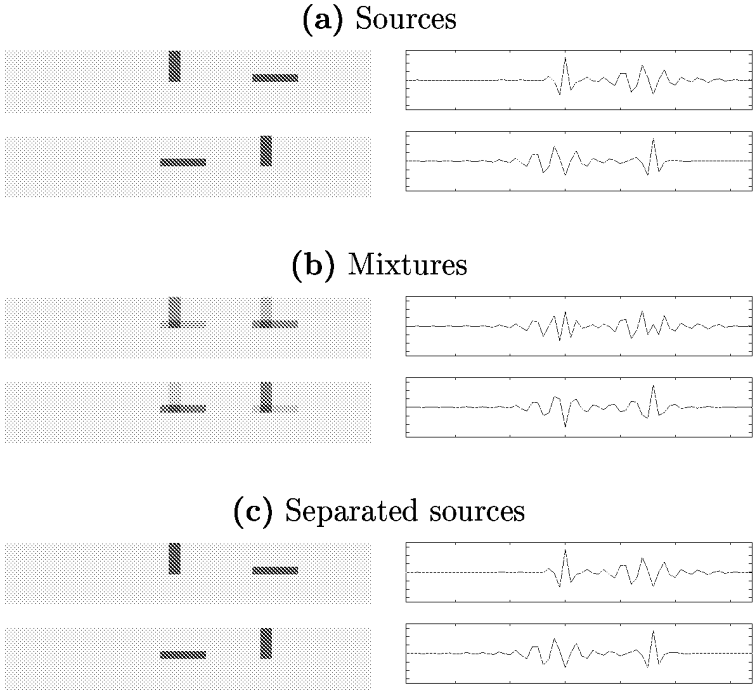

Figure 7: Sources, mixtures, and reconstructed sources in both time-frequency phase plane (left) and time domain (right).

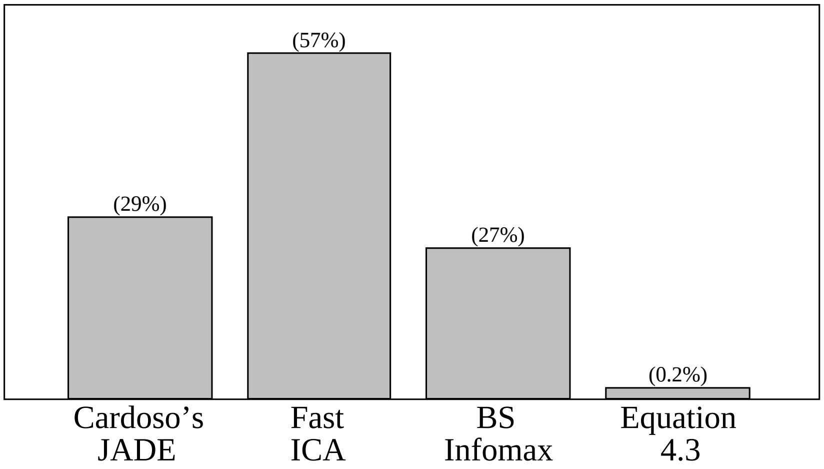

Figure 8: Percentage relative error of separation of the artificial sparse sources recovered by JADE, fast ICA, Bell-Sejnowski Infomax, and equation 4.3.

where the term $\sum_k h(c_k)$ may be considered a penalty for nonsparsity. 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,

$$\|w\|_2^2 \geq 1. \tag{5.2}$$

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

$$\|(I - P^{j-1})w^j\|_2^2 \geq 1\ , \tag{5.3}$$

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

$$c = c^+ - c^-, \quad c^+ \geq 0, \quad c^- \geq 0$$

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

which transforms equations 5.1 and 5.3 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 Nonovercomplete Dictionaries

In important applications (Tang, Pearlmutter & Zibulevsky, 2000; Tang, Pearlmutter, Zibulevsky, Hely, & Weisend, 2000; Tang, Phung, Pearlmutter, & Christner, 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 that compromises between statistical and computational efficiency. In our experience, this approach provides a high quality of separation in a reasonable amount of time.

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

Let us denote the dual basis,

$$\Psi = \Phi^{-1}, \tag{6.1}$$

and suppose that coefficients of decomposition of the sources,

$$C = S\Psi, \tag{6.2}$$

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

Let Y be the decomposition of the sensor signals,

$$Y = X\Psi. \tag{6.3}$$

Multiplying both sides of equation 1.3 by Ψ from the right and taking into account equations 6.2 and 6.3, we obtain

$$Y = AC + \zeta, \tag{6.4}$$

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 sections 3.1 to 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 3.9 and 3.14 are transformed into

$$\min_{A,C} \frac{1}{2\sigma^2}\|AC - Y\|_F^2 + \sum_{j,k} \beta_j h(C_{jk}) \tag{6.5}$$

and

$$\min_{A,C} \frac{1}{2\sigma^2}\|AC - Y\|_F^2 + \sum_{j} \frac{2\sum_k |C_{jk}|}{\sqrt{K^{-1}\sum_k C_{jk}^2}}. \tag{6.6}$$

The objective, equation 4.3, becomes

$$\min_{W,C} -K \log|\det W| + \frac{1}{2}\|C - WY\|_F^2 + \mu \sum_{j,k} \beta_j h(C_{jk}). \tag{6.7}$$

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

$$\min_{W} -K \log|\det W| + \sum_{j,k} \beta_j h((WY)_{jk}). \tag{6.8}$$

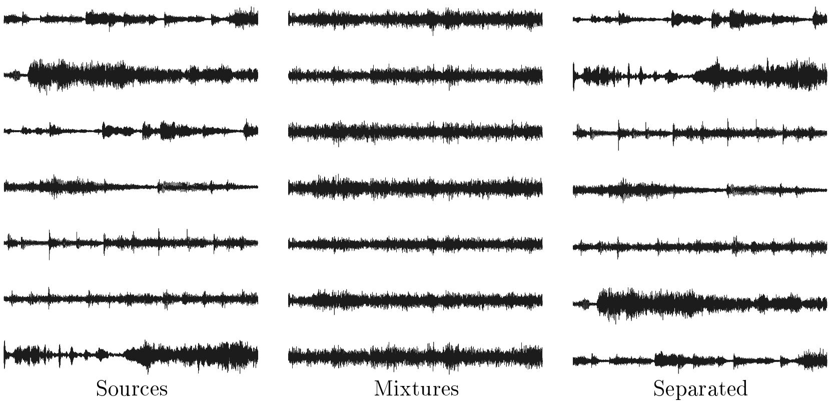

Figure 9: Separation of musical recordings taken from commercial digital audio CDs (5 second fragments).

Also other known methods (e.g., Lee et al., 1999; Lewicki & Sejnowski, in press), 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.

6.1 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 40,000 samples after downsampling by a factor of 5 (see 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 (see Figure 10, left).

In this case Y (see equation 6.3) 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 (see equation 6.8) with $\beta_j = 1$ and $h_\lambda(\cdot)$ defined by equations 3.3 and 3.4 with the parameter $\lambda = 10^{-4}$. Unconstrained minimization was performed by a BFGS quasi-Newton algorithm (MATLAB function FMINU.)

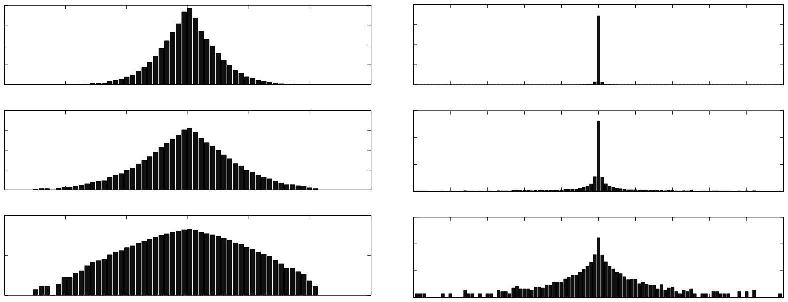

Figure 10: Histogram of sound source values (left) and spectrogram coefficients (right), shown with linear y-scale (top), square root y-scale (center), and logarithmic y-scale (bottom).

This algorithm separated the sources with a relative error of 0.67% for the least well-separated source (error computed according to equation 3.15). We also applied the BS Infomax algorithm (Bell & 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 minutes for BFGS to 5 minutes for BS Infomax).

For comparison we tested the JADE (Cardoso, 1999a, 1999b), FastICA (Hyvärinen, 1998, 1999), and BS Infomax algorithms on the raw sensor signals. Resulting relative errors (see 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 (see equation 3.8). 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 (in press) and Lewicki and Olshausen (1998) or by variational methods (Jordan, Ghahramani, Jaakkola, & Saul, 1999). It would be interesting to compare these possibilities to the other methods presented in this article.

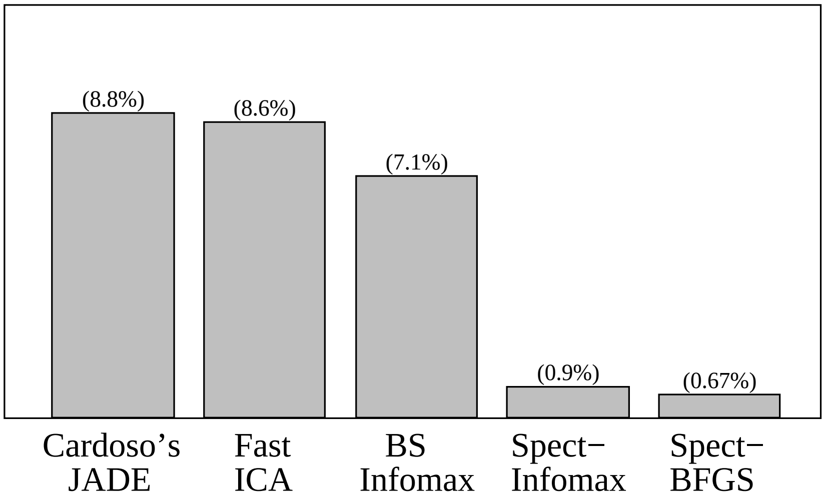

Figure 11: Percentage relative error of separation of seven musical sources recovered by JADE; fast ICA; Bell-Sejnowski Infomax; Infomax, applied to the spectrogram coefficients; and BFGS minimization of the objective (see equation 6.8) with the spectrogram coefficients.

Another important direction is toward the problem of simultaneous blind deconvolution and separation, as in Lambert (1996). In this case, the matrices *A* and *W* will have linear filters as an elements, and multiplication by an element corresponds to convolution. Even in this matrix-of-filters context, most of the formulas 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 nonconvex quadratic constraints. Finally, much faster solutions may be obtained by using nonovercomplete 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., & Yang, H. H. (1996). A new learning algorithm for blind signal separation. In D. Touretzky, M. Mozer, & M. Hasselmo (Eds.), *Advances in neural information processing systems, 8*. Cambridge, MA: MIT Press.

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

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

Belouchrani, A., & 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* (pp. 49–53). Las Vegas, NV.

Ben-Tal, A., & 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., & Scargle, J. (1995). *About wavelab* (Tech. Report). Stanford, CA: Department of Statistics, Stanford University. Available online at: 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. Available online at: http://sig.enst.fr/~cardoso/guidesepsou.html.

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

Chen, S. S., Donoho, D. L., Saunders, M. A., Johnstone, I., & Scargle, J. (1995). About atomizer (Tech. Rep.). Stanford, CA: Department of Statistics, Stanford University. Available online at: http://www-stat.stanford.edu/~donoho/Reports/.

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

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

Hyvärinen, A. (1998). The Fast-ICA MATLAB package. Available online at: 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.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. In M. I. Jordan (Ed.), *Learning in graphical models* (pp. 105–161). Norwell, MA: Kluwer.

Lambert, R. H. (1996). *Multichannel blind deconvolution: FIR matrix algebra and separation of multipath mixtures*. Unpublished doctoral dissertation, University of Southern California.

Lee, T. W., Lewicki, M. S., Girolami, M., & Sejnowski, T. J. (1999). Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letters*, *6*, 87–90.

Lewicki, M. S., & Olshausen, B. A. (1998). A probabilistic framework for the adaptation and comparison of image codes. *Journal of the Optical Society of America A: Optics, Image Science, and Vision, 16*, 1587–1601.

Lewicki, M. S., and Sejnowski, T. J. (in press). Learning overcomplete representations. *Neural Computation*.

Makeig, S. (1999). ICA/EEG toolbox. San Diego, CA: Computational Neurobiology Laboratory, Salk Institute. Available online at: http://www.cnl.salk.edu/∼tewon/ica_cnl.html.

Mallat, S. (1998). *A wavelet tour of signal processing*. Orlando, FL: Academic Press.

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

Pajunen, P., Hyvdrinen, A., & Karhunen, J. (1996). Non-linear blind source separation by self-organizing maps. In *Proceedings of the International Conference on Neural Information Processing*. Berlin: Springer-Verlag.

Pearlmutter, B. A., and Parra, L. C. (1996). A context-sensitive generalization of ICA. In *Proceedings of the International Conference on Neural Information Processing* (pp. 151–157). Berlin: Springer-Verlag.

Rowe, D. B. (1999). *Bayesian blind source separation*. Submitted.

Tang, A. C., Pearlmutter, B. A., & Zibulevsky, M. (2000). Blind separation of neuromagnetic responses. *Neurocomputing, 32–33* (Special Issue), 1115–1120.

Tang, A. C., Pearlmutter, B. A., Zibulevsky, M., Hely, T. A., & Weisend, M. P. (2000). An MEG study of response latency and variability in the human visual system during a visual-motor integration task. In S. A. Solla, T. K. Leen, & K. R. Müller (Eds.), *Advances in neural information processing systems, 12* (pp. 185–191). Cambridge, MA: MIT Press.

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

Tseng, P. (2000). *Convergence of block coordinate descent method for nondifferentiable minimization*. Submitted.

Zibulevsky, M., Pearlmutter, B. A., Bofill, P., & Kisilev, P. (in press). Blind source separation by sparse decomposition in a signal dictionary. In S. J. Roberts & R. M. Everson (Eds.), *Independent components analysis: Princeiples and practice*. Cambridge: Cambridge University Press.

Received July 28, 1999; accepted July 11, 2000.