BLIND SEPARATION OF SOURCES WITH SPARSE REPRESENTATIONS IN A GIVEN SIGNAL DICTIONARY

Michael Zibulevsky *Barak A. Pearlmutter*

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

ABSTRACT

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

1. INTRODUCTION

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

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

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

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

The scalar functions $\varphi_k(t)$ are called *atoms* or *elements* of the dictionary. These elements do not have to be linearly independent, and instead may form an overcomplete dictionary. Important examples are wavelet-related dictionaries (wavelet packets, stationary wavelets, *etc.*, see for example [1, 2] and references therein), or learned dictionaries [3, 4, 5]. Sparsity means that only a small number of the coefficients C_{ik} differ significantly from zero.

Supported in part 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.

We consider 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) and (2) then take the following simple form

$$X = AS + \xi \tag{3}$$
$$S = C\Phi \tag{4}$$

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

$$X \approx AC\Phi$$

Our goal therefore can be formulated as follows:

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

2. PROBABILISTIC FRAMEWORK

In order to derive a maximum *a posteriori* solution, we consider the blind source separation problem in a probabilistic framework [6, 7]. Suppose that the coefficients C_{ik} in source decomposition (4) are statistically independent random variables with a probability density function (pdf) of an exponential type

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

This kind of distribution is widely used for modeling sparsity [3, 5]. A reasonable choice of $h(c)$ may be

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

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

$$\begin{aligned} h_1(c) &= |c| - \log(1 + |c|) && (7)\\ h_\lambda(c) &= \lambda h_1(c/\lambda) && (8) \end{aligned}$$

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

We also suppose *a priori* that the mixing matrix A is uniformly distributed over the range of interest, and that the noise $\xi(t)$ in (3) is a spatially and temporally uncorrelated Gaussian process[1] with zero mean and variance σ^2.

2.1. Maximum a posteriori approach

We wish to maximize the posterior probability

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

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

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

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

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

If the prior pdf $P(A)$ is uniform, it can be dropped[2] from (9). In this way we are left with the problem

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

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

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

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

[1] The iid noise assumption is for simplicity of exposition and can be easily removed.

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

One can consider this objective as a generalization of [5] by incorporating the matrix Φ, or as a generalization of [1] 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,

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

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

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

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

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

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

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

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

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

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

2.2. 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 [8, 2] are shown in Figure 1. We used the ATOMIZER [9] and WAVELAB [10] MATLAB packages for fast multiplication by Φ and Φ^T. We created

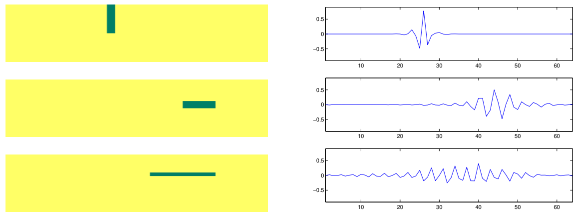

Figure 1: Examples of atoms: time-frequency phase plane (left) and time plot (right.)

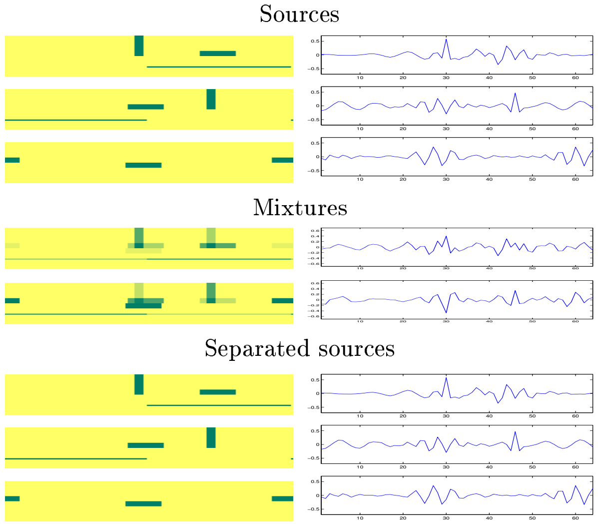

Figure 2: Sources, mixtures and reconstructed sources, in both time-frequency phase plane (left) and time domain (right).

three very sparse sources (Figure 2, top), each composed of only two or three atoms. The first two sources have significant cross-correlation, equal to 0.34, which makes separation difficult for conventional methods. Two synthetic sensor signals (Figure 2, center) were obtained as a linear mixture of the sources. In order to measure the accuracy of the separation, we normalized the original sources with $\|S_j\|_2 = 1$, and the estimated sources with $\|\widetilde{S}_j\|_2 = 1$. The error was then computed as

$$\text{Error} = \frac{\|\widetilde{S}_j - S_j\|_2}{\|S_j\|_2} \cdot 100\% \qquad (19)$$

We tested two methods with this data. The first method used the objective function (13) and the constraints (15), while the second method used the objective function (18). As a tool for constrained optimization we used the PBM method [11]. Unconstrained optimization was produced by the method of conjugate gradients using the TOMLAB package [12]. The same tool was used for internal unconstrained optimization in PBM.

In all the cases we used $h_\lambda(\cdot)$ defined by (7) and (8), with the parameter $\lambda = 0.01$. Another parameter $\sigma^2 = 0.0001$. The resulting errors of the source estimates were 0.09% and 0.02% by the first and the second method respectively. The estimated sources are shown in the bottom three traces of Figure 2. They are visually indistinguishable from the original sources, shown in top three traces of Figure 2.

It is important to note 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 from actual solution (in order to get such an initial guess, one can use a clustering-type algorithm, as in [13]).

Second, the method of conjugate gradients requires a few thousand iterations to converge, which takes about 5 min at Pentium 300 MHz processor even for this very small problem[3]. In the remaning part of the paper we present few other approaches, which help to stabilize and accelerate optimization.

3. EQUAL NUMBER OF SOURCES AND SENSORS: MORE ROBUST FORMULATIONS

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

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

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

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

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

It can be shown, that in the noiseless case, $\sigma \approx 0$, the problem (20)–(21) is equivalent to the maximum a

[3]Our preliminary experiments with other algorithms (like truncated Newton method) give a hope to reduce this timing by an order of magnitude or more.

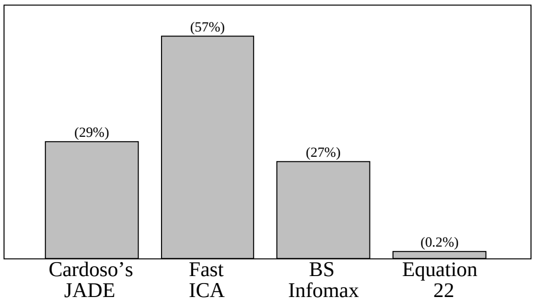

Figure 3: Percent relative error of separation of the artificial sparse sources recovered by (1) JADE, (2) Fast ICA, (3) Bell-Sejnowski Infomax, (4) Equation 22.

posteriori formulation (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

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

When the noise is zero and Φ is the identity matrix, we can substitute $C = WX$ and obtain the Bell-Sejnowski Infomax objective [14]

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

Experiment: equal numbers of sources and sensors

We created two sparse sources with strong cross-correlation of 0.52. Separation, produced by minimization of the objective function (22), gave an error of 0.23%. Robust convergense was achieved when we started from random uniformely distributed points in C and W.

For comparison we tested the JADE [15, 16], FastICA [17, 18] and Bell-Sejnowski Infomax [14, 19] algorithms on the same signals. All three codes were obtained from the refereed websites and were used with default setting of all parameters. The resulting relative errors (Figure 3) 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 Pentium 300 MHz). For comparision, JADE, FastICA and Infomax take only few seconds. Below we will consider some options for acceleration.

4. FAST SOLUTION IN NON-OVERCOMPLETE DICTIONARIES

In important applications, 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 [3, 4, 5].

Let us denote the dual basis

$$\Psi = \Phi^{-1} \tag{24}$$

and suppose that coefficients of decomposition of the sources

$$C = S\Psi \tag{25}$$

are sparse and statistically 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

$$Y = X\Psi \tag{26}$$

Multiplying both sides of (3) by Ψ from the right and taking into account (25) and (26), we obtain

$$Y = AC + \zeta \; , \tag{27}$$

where ζ is decomposition of the noise

$$\zeta = \xi\Psi \; . \tag{28}$$

Here we consider an "easy" situation, when ζ is a white noise, that requires orthogonality of Ψ. We can see that all the objective functions from the previous sections remain valid if we remove from them Φ (substituting instead the identity matrix) and replace the sensor signal X by its decomposition Y. For example, maximum *a posteriori* objectives (13) and (18) are transformed into

$$\min_{A,C} \; \frac{1}{2\sigma^2}\|AC - Y\|_F^2 + \sum_{j,k} \beta_j h(C_{jk}) \tag{29}$$

and

$$\min_{A,C} \; \frac{1}{2\sigma^2}\|AC - Y\|_F^2 + \sum_j \frac{2\sum_k |C_{jk}|}{\sqrt{K^{-1}\sum_k C_{jk}^2}} \tag{30}$$

The objective (22) becomes

$$\min_{W,C} \; -K \log|\det W| + \frac{1}{2}\|C - WY\|_F^2 + \mu \sum_{j,k} \beta_j h(C_{jk}) \tag{31}$$

In this case we can further assume that the noise is zero. substitute $C = WY$ and obtain the Bell-Sejnowski Infomax objective [14]

$$\min_{W} \; -K \log|\det W| + \sum_{j,k} \beta_j h((WY)_{jk}) \tag{32}$$

Also other known methods (for example, [20, 3]), 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 than the raw signals had originally. Also, statistical independence of the coefficients is a more reasonable assumption than statistical 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.

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 4, right) is much higher then the sparsity of the original signal (Figure 4, left)

In this case Y (26) 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 (32) with $\beta_j = 1$ and $h_\lambda(\cdot)$ defined by (7),(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 (19).) We also applied the Bell-Sejnowski Infomax algorithm [14] implemented in [19] to the spectrogram coefficients Y of the sensor signals. Separation errors were slightly larger: 0.9%, but the computing time was improved (from 30 min for BFGS to 5 min for Infomax).

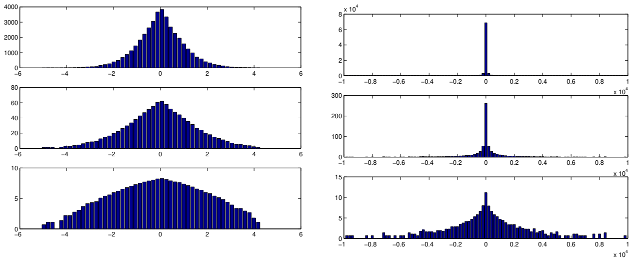

Figure 4: Histogram of sound source values (left) and spectrogram coefficients (right), shown with linear y-scale (top), square root y-scale (center) and logarithmic y-scale (bottom).

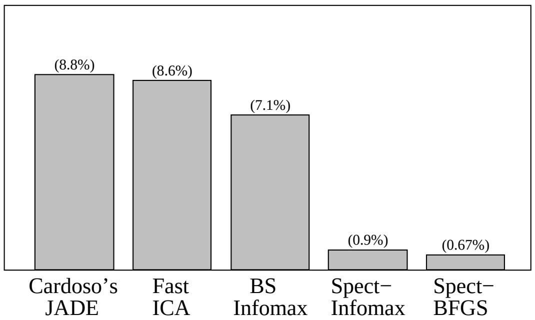

Figure 5: Percent relative error of separation of seven musical sources. recovered by (1) JADE, (2) Fast ICA, (3) Bell-Sejnowski Infomax, (4) Infomax, applied to the spectrogram coefficients, (5) BFGS minimization of the objective (32) with the spectrogram coefficients.

For comparison we tested JADE [15, 16], FastIca [17, 18] and Infomax algorithms on the raw sensor signals. Resulting relative errors (Figure 5) confirm the significant (by a factor of more than 10) superiority of the sparse decomposition approach.

The method described in this section, that combines spectrogram transformations with the Infomax algorithm, is included by Scott Makeig into the ICA/EEG toolbox [19].

5. CONCLUSIONS

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

6. REFERENCES

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

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

[3] M. Lewicki and T. Sejnowski, "Learning overcomplete representations," *Neural Computation*, 1998. to appear.

[4] M. Lewicki and B. Olshausen, "A probabilistic framework for the adaptation and comparison of image codes," *Journal of the Optical Society of America*, 1999. in press.

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

[6] A. Belouchrani and J.-F. Cardoso, "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*, (Las Vegas, NV), pp. 49–53, Dec. 1995. In press.

[7] B. A. Pearlmutter and L. C. Parra, "A context-sensitive generalization of ICA," in *International Conference on Neural Information Processing*, (Hong Kong), pp. 151–157, Springer-Verlag, Sept. 24–27 1996.

[8] R. Coifman and M. Wickerhauser, "Entropy-based algorithms for best-basis selection," *IEEE Transactions on Information Theory*, vol. 38, pp. 713–718, 1992.

[9] S. S. Chen, D. L. Donoho, M. A. Saunders, I. Johnstone, and J. Scargle, "About atomizer," tech. rep., Department of Statistics, Stanford University, 1995. http://www-stat.stanford.edu/~donoho/Reports/.

[10] J. Buckheit, S. Chen, D. Donoho, I. Johnstone, and J. Scargle, "About wavelab," tech. rep., Department of Statistics, Stanford University, 1995. http://www-stat.stanford.edu/~donoho/Reports/.

[11] A. Ben-Tal and M. Zibulevsky, "Penalty/barrier multiplier methods for convex programming problems," *SIAM Journal on Optimization*, vol. 7, no. 2, pp. 347–366, 1997.

[12] K. Holmstrom and M. Bjorkman, "The TOMLAB NLPLIB," *Advanced Modeling and Optimization*, vol. 1, pp. 70–86, 1999. http://www.ima.mdh.se/tom/.

[13] P. Bofill and M. Zibulevsky, "Sparse underdetermined ICA: Estimating the mixing matrix and the sources separately," Tech. Rep. UPC-DAC-2000-7, Universitat Politecnica de Catalunya, 2000. http://www.ac.upc.es/homes/pau/sounds.html.

[14] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[15] J.-F. Cardoso, "High-order contrasts for independent component analysis," *Neural Computation*, vol. 11, no. 1, pp. 157–192, 1999.

[16] J.-F. Cardoso, "JADE for real-valued data," 1999. http://sig.enst.fr:80/~cardoso/guidesepsou.html.

[17] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[18] A. Hyvärinen, "The Fast-ICA MATLAB package," 1998. http://www.cis.hut.fi/~aapo/.

[19] S. Makeig, "ICA/EEG toolbox." Computational Neurobiology Laboratory, the Salk Institute, 1999. http://www.cnl.salk.edu/~tewon/ica_cnl.html.

[20] T. Lee, M. Lewicki, M. Girolami, and T. Sejnowski, "Blind source separation of more sources than mixtures using overcomplete representations," *IEEE Sig. Proc. Lett.*, 1998. to appear.