Independent Component Analysis in the Presence of Gaussian Noise by Maximizing Joint Likelihood

Aapo Hyvärinen

Helsinki University of Technology, Lab. of Computer and Information Science, P.O. Box 2200, FIN-02015 HUT, Finland

aapo.hyvarinen@hut.fi

To appear in Neurocomputing, 1998(?)

We consider the estimation of the data model of independent component analysis when gaussian noise is present. We show that the joint maximum likelihood estimation of the independent components and the mixing matrix leads to an objective function already proposed by Olshausen and Field using a different derivation. Due to the complicated nature of the objective function, we introduce approximations that greatly simplify the optimization problem. We show that the presence of noise implies that the relation between the observed data and the estimates of the independent components is non-linear, and show how to approximate this non-linearity. In particular, the non-linearity may be approximated by a simple shrinkage operation in the case of supergaussian (sparse) data. Using these approximations, we propose an efficient algorithm for approximate maximization of the likelihood. In the case of supergaussian components, this may be approximated by simple competitive learning, and in the case of subgaussian components, by anti-competitive learning.

Key words: Independent component analysis, blind source separation, maximum likelihood, competitive learning, neural networks.

1 Introduction

Independent Component Analysis (ICA) [1,4,6,7,11,13,15,18] is a statistical technique whose goal is to represent a set of random variables as linear combinations of statistically independent component variables. Important applications of ICA are in blind source separation [13] and feature extraction [2,14]. One may formulate ICA as the estimation of the following linear generative model for the data:

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n} \tag{1}$$

where $\mathbf{x} = (x_1, x_2, ..., x_m)$ is the vector of observed random variables, $\mathbf{s} = (s_1, s_2, ..., s_n)$ is the vector of the latent variables called the independent components, and $\mathbf{A}$ is an unknown constant matrix, called the mixing matrix. The vector $\mathbf{n}$ is noise, and is often omitted. The fundamental assumption is that the independent components s_i are mutually *statistically independent*. The basic problem of ICA is then to estimate both the mixing matrix $\mathbf{A}$ and the realizations of the independent components s_i *using only observations of the mixtures* x_j. In the maximum likelihood framework used in this paper, it is also assumed that we know, at least approximately, the probability distributions of the independent components. The fundamental restriction of the model is that the independent components (expect perhaps one) must be non-Gaussian. It is also assumed that the s_i have zero mean but this is in fact no restriction, as this can always be accomplished by subtracting the mean from the random vector $\mathbf{x}$. Note that the independent components and the columns of $\mathbf{A}$ can only be estimated up to a multiplicative constant, because any constant multiplying an independent component in eq. (1) could be cancelled by dividing the corresponding column of the mixing matrix $\mathbf{A}$ by the same constant. For mathematical convenience, one usually defines that the independent components s_i have unit variance [7].

In this paper, we approach the noisy ICA problem using maximum likelihood estimation. We estimate jointly $\mathbf{A}$ and the $\mathbf{s}$ in model (1), which leads to an objective function that was already proposed by Olshausen and Field [19] for sparse coding, using a different derivation (Section 2). This approach gives certain important results. First, the presence of noise in (1) implies that the optimal reconstruction of the independent components is *not* a linear function of the $\mathbf{x}$, as in the noiseless case [7]. We derive closed-form expressions for this non-linear reconstruction using certain assumptions and approximations, and show that often a good approximation is provided by simple 'shrinkage' or 'truncation' operators (Section 3). Second, we show how the optimization of the objective function is greatly simplified by these approximations, and propose an efficient and simple algorithm performing the joint estimation (Section 4). As a modification of this algorithm, it is shown that for supergaussian (sparse) components, a simple variation of the k-means algorithm is enough to estimate the mixing matrix (Section 5). Furthermore, this competitive algorithm allows us to relax the conventional assumption in ICA estimation that the number of the independent component s_i is not larger than the dimension of the observed data, i.e., $n \leq m$, without resorting to computationally complex methods. The corresponding modification of the algorithm for subgaussian components uses an anti-competitive mechanism. Applications of the introduced estimation methods to denoising are discussed in Section 6. Finally, simulation results are presented in Section 7, and the results are discussed and compared to previous work in Section 8.

2 Maximum Likelihood Estimation for ICA

The estimation of $\mathbf{A}$ and $\mathbf{s}$ in (1) can be accomplished by a maximum likelihood estimation. Assume that we have observed T vectors $\mathbf{x}(t), t = 1, ..., T$ generated according to (1). One often knows, at least approximatively, the densities $p_i(.)$ of the s_i, as used in ordinary (noise-free) maximum likelihood estimation of $\mathbf{A}$ [21,1]. Denote the corresponding (negative) log-densities (or log-likelihoods) by $f_i(.) = -\log p_i(.)$. For simplicity, we may assume that the s_i all have identical densities, in which case the (negative) log-density of $\mathbf{s}$ is by independence of the form $\sum_{i=1}^{n} f(.)$. The noise $\mathbf{n}$ in the model is assumed to be Gaussian with known covariance matrix Σ. (Throughout this paper, the noise covariance is assumed to be known.) Thus one obtains the log-likelihood:

$$\log L(\mathbf{A}, \mathbf{s}(1), ..., \mathbf{s}(T)) = \\ -\sum_{t=1}^{T} \left[\frac{1}{2} \|\mathbf{A}\mathbf{s}(t) - \mathbf{x}(t)\|^2_{\Sigma^{-1}} + \sum_{i=1}^{n} f(s_i(t)) \right] + C \tag{2}$$

where $\|\mathbf{e}\|^2_{\Sigma^{-1}}$ is defined as $\mathbf{e}^T \Sigma^{-1} \mathbf{e}$, the $\mathbf{s}(t)$ are the realizations of the independent components, and C is an irrelevant costant. Recall that the s_i are here constrained to have unit variance. This likelihood is essentially the same as the objective function proposed by Olshausen and Field [19] as an approximation of the likelihood of $\mathbf{A}$. As the above derivation shows, however, it is not necessary to consider this objective function as an approximation: for joint estimation of $\mathbf{A}$ and the $\mathbf{s}(t)$, it is the exact likelihood. The reason for this difference is that Olshausen and Field considered the $\mathbf{s}(t)$ to be nuisance parameters that should be integrated out. However, in many applications of ICA, like blind source separation, the $\mathbf{s}(t)$ may be even more interesting than $\mathbf{A}$.

The maximization of (2) is not an easy task, because it is a function of both the matrix $\mathbf{A}$ and the values $\mathbf{s}(t), t = 1, ..., T$, which makes $n \times m + n \times T$ variables. Therefore, we shall introduce in the following section approximations that considerably reduce the dimension of the problem.

3 Non-linear Reconstruction of the Independent Components

A crucial difference between the noisy and noiseless ICA models is that in the presence of noise the maximization of the likelihood leads to non-linear reconstruction of the independent components, as was noted in [19]. In other words, the estimate of $\mathbf{s}(t)$ is not obtained simply by $\hat{\mathbf{s}}(t) = \hat{\mathbf{A}}^{-1}\mathbf{x}(t)$, where $\hat{\mathbf{s}}(t)$ and $\hat{\mathbf{A}}$ are the corresponding estimates of $\mathbf{A}$ and $\mathbf{s}(t)$. The non-linear

relation is, in general, very complex. Therefore, we give in this section two closed-form approximations for this non-linear reconstruction. In particular, we show that if the independent components are supergaussian (or sparse), this non-linear reconstruction may be approximated by a shrinkage of the $s_i(t)$ towards zero. This operation is formally very similar to the wavelet shrinkage method by Donoho et al [8].

In this section, we assume for simplicity that the dimension of the data m equals the number of independent components n. The results can be easily extended for the case $m > n$, but not for the case $m < n$. Taking the gradient of the Lagrangian of the log-likelihood (2) with respect to the $\mathbf{s}(t), t = 1, ..., T$ and equating this to 0, one obtains at the optimum (see Appendix)

$$\hat{\mathbf{A}}^T \Sigma^{-1} \hat{\mathbf{A}} \hat{\mathbf{s}}(t) - \hat{\mathbf{A}}^T \Sigma^{-1} \mathbf{x}(t) + f'(\hat{\mathbf{s}}(t)) = \Lambda \hat{\mathbf{s}}(t) \tag{3}$$

where the derivative of the log-density, f', is applied separately on each component of the vector $\hat{\mathbf{s}}(t)$ (this convention will be used throughout the paper). The diagonal matrix Λ of the Lagrangian coefficients can be easily evaluated to give at the optimum approximately $\Lambda = E\{f'(s_i)s_i\}\mathbf{I} = \mathbf{I}$ due to the Bussgang equality (assuming that $-f$ is the true log-likelihood of the data) .

Equation (3) can be used to derive two simple approximations of the optimal non-linearity, using two different assumptions. First, assume that the noise level is small and that the derivative of the log-density f' is a smooth function. Then a first-order approximation (with respect to noise level) of this equation may be obtained by replacing $\hat{\mathbf{s}}(t)$ in the last two terms by $\hat{\mathbf{A}}^{-1}\mathbf{x}(t)$. This can be solved for $\hat{\mathbf{s}}(t)$, giving

$$\hat{\mathbf{s}} = \hat{\mathbf{A}}^{-1}\mathbf{x} - \hat{\mathbf{A}}^{-1}\Sigma\hat{\mathbf{A}}^{-T}(f'(\hat{\mathbf{A}}^{-1}\mathbf{x}) - \hat{\mathbf{A}}^{-1}\mathbf{x}) \tag{4}$$

where the index t has been dropped for simplicity. Alternatively, we may assume that the covariance matrix has a particularly simple structure: $\Sigma = \sigma^2\mathbf{A}\mathbf{A}^T$. Then (3) gives

$$\hat{\mathbf{s}} = h(\hat{\mathbf{A}}^{-1}\mathbf{x}) \tag{5}$$

where the scalar component-wise function h is obtained by inverting the relation (see Appendix)

$$h^{-1}(u) = (1 - \sigma^2)u + \sigma^2 f'(u). \tag{6}$$

The assumption on covariance means that the noise can be considered to be added to the independent components before the mixing (but this need not be

the case physically). Moreover, if the mixing matrix is orthogonal (as is often approximately the case in feature extraction), this assumption is valid for the classical noise covariance of $\sigma^2\mathbf{I}$. Note that if the constraint of unit variance is discarded, σ^2 inside the parentheses in (6) disappears; this equation is used extensively in the method of sparse code shrinkage, see Section 6.

The two approximations in (4) and (5) simplify the optimization of (2) considerably: substituting the right-hand side of either (4) or (5) for $\mathbf{s}$ in (2), one has reduced the dimension of the problem to the dimension of the matrix $\mathbf{A}$, instead of the original, rather intractable dimension $n \times m + n \times T$. Whether it is better to use (4) or (5) depends basically on which assumptions are more correct. If f' is smooth and noise covariance is far from the form $\sigma^2\mathbf{A}\mathbf{A}^T$, (4) may be better; otherwise, (5) will probably give better results.

As an important example, let us assume that the density of the independent components is *double exponential* (or Laplace), which is a classical example of a supergaussian (or sparse) distribution, i.e., a distribution of positive kurtosis [11]. supergaussian distribution have typically heavier tails than the Gaussian distribution, and a peak at zero. Such distributions are found, for example, in many situations of feature extraction and speech processing. Then one has $f(u) = \sqrt{2}|u|$ (plus an irrelevant constant), and one sees easily that for $\sigma^2 < 1$, the function h in (5) has the form of a *shrinkage* operator, as depicted in Fig. 1:

$$h(u) = \frac{1}{1-\sigma^2}\text{sign}(u)\max(0, |u| - \sqrt{2}\sigma^2) \tag{7}$$

(To see this, approximate f' by a sequence of continuous functions and take the limit.) This means that h first decreases the absolute value of its argument by a certain amount (hence 'shrinkage'), and then performs a simple rescaling. Intuitively, such a shrinkage has appealing properties: it suppresses values of independent components that are very small, thus reducing noise. Indeed, it can also be motivated as an optimal de-noising method according to minimax estimation theory [8].

Another interesting example is found when the s_i have a *uniform* distribution. The uniform distribution is a typical example of a subgaussian distribution, i.e., a distribution of negative kurtosis [11]. subgaussian distributions are typically flatter and have lighter tails when compared to the Gaussian distribution. Then one obtains the *truncation* operator $h(u) = \text{sign}(u)\min(|u|, \sqrt{3})$. This is also intuitively appealing: since an uniform variable of unit variance cannot have values that are outside the interval $[-\sqrt{3}, \sqrt{3}]$, all values must be forced to stay in that interval.

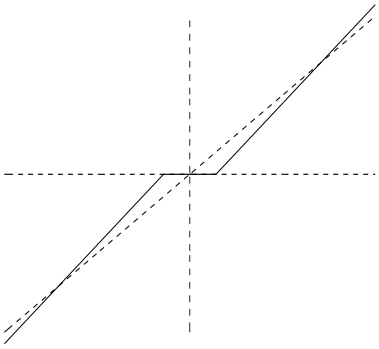

Fig. 1. Plot of the (scaled) shrinkage function. The effect of the function is to reduce the absolute value of its argument by a certain amount, which depends on the parameters, and then rescale. Small arguments are set to zero. This reduces Gaussian noise for supergaussian (sparse) independent components.

4 Optimization by alternating variables method

Using the approximations introduced in Section 3, the optimization of (2) can be accomplished by a simple alternating variables method, which has already been used in similar estimation tasks [20]. The method is based on first optimizing the objective function with respect to $\mathbf{A}$ for fixed $\mathbf{s}(t)$, then optimizing with respect to the $\mathbf{s}(t)$ for fixed $\mathbf{A}$, and so on. The optimization with respect to $\mathbf{A}$ for fixed $\mathbf{s}(t)$ is evidently accomplished by a simple least-squares fit, and the optimization with respect to $\mathbf{s}(t)$ using the methods of the preceding section. Thus, the algorithm has the following form:

(i) Take some initial value for $\hat{\mathbf{A}}_0$. Let $k = 1$.
(ii) Compute the $\hat{\mathbf{s}}_k(t)$ by either (4) or (5), using $\hat{\mathbf{A}}_{k-1}$ as the estimate of $\mathbf{A}$. Normalize the components of $\hat{\mathbf{s}}_k(t)$ to have unit variance.
(iii) Update $\hat{\mathbf{A}}_k = E\{\mathbf{x}\hat{\mathbf{s}}_k^T\}[E\{\hat{\mathbf{s}}_k\hat{\mathbf{s}}_k^T\}]^{-1}$
(iv) Increment k and go back to step 2 if not converged.

This algorithm may be further simplified as follows. First, the matrix in brackets in step 3 is usually close to identity (especially near the solution), so it may be discarded. Second, the inversion of the matrix $\hat{\mathbf{A}}$ in (4) and (5) may be avoided by first sphering (or whitening) the data. Sphering means that the covariance matrix of $\mathbf{x}$ is made equal to unity, i.e., $E\{\mathbf{x}\mathbf{x}^T\} = \mathbf{I}$, which is possible by a simple linear transformation [7]. Then it is easy to see that one has $\mathbf{A}^{-1} = \mathbf{A}^T(\mathbf{I} - \Sigma)^{-1}$, which can be simplified in the case $\Sigma = \sigma^2\mathbf{A}\mathbf{A}^T$ to yield $\mathbf{A}^{-1} = \mathbf{A}^T/(1 - \sigma^2)$. Thus, $\hat{\mathbf{A}}^{-1}$ may be replaced by $\hat{\mathbf{A}}^T(\mathbf{I} - \Sigma)^{-1}$ when equations (4) and (5) are computed in the algorithm. After these two simplifications, no matrix inversions are needed in the algorithm. Moreover, the stability of the algorithm can then be improved by orthogonalizing the matrix $\hat{\mathbf{A}}$ so that $\hat{\mathbf{A}}\hat{\mathbf{A}}^T = \mathbf{I} - \Sigma$ after every iteration.

Methods for the neural implementation of sphering can be found, e.g. in [6,15]. Note that if the data is sphered and (5) is used, our algorithm becomes somewhat similar to the algorithm proposed in [16] on a more heuristic basis.

5 Competitive and Anti-Competitive mechanisms

5.1 The dichotomy in maximum-likelihood estimation

It is a well-known fact that the conventional maximum-likelihood (or infomax) estimation of the mixing matrix is very robust with respect to misspecification of the densities of the independent components [1,21]. Indeed, it has been proven in a similar context [12] that for any distribution of the s_i, and for any well-behaving non-quadratic function g one can always estimate the mixing matrix by using either g or $-g$ as an approximation of the log-likelihood. Therefore, it seems very likely that for estimation of $\mathbf{A}$, it would be enough to use just two different densities in (2). We take two 'densities', one which corresponds to a distribution that is extremely supergaussian (i.e., has a very large positive kurtosis), and another one which corresponds to an extremely subgaussian distibution. (i.e., one which has an extremely negative kurtosis). It seems then reasonable to assume that whatever the distributions of the s_i may be (as long as they are identical for all independent components), one of these learning rules performs the estimation of $\mathbf{A}$ correctly. The estimation of the $\mathbf{s}(t)$ can then be performed afterwards, by minimizing (2) separately with respect to $\mathbf{s}(t)$ for every t. These extremely supergaussian and subgaussian 'densities' lead to competitive and anti-competitive learning, respectively, when used in connection with the algorithm of the preceding section.

5.2 A competitive mechanism for supergaussian components

To estimate $\mathbf{A}$ in the ICA model when the independent components are *supergaussian* (or sparse, or have positive kurtosis), one may thus use a log-density f that is extremely supergaussian. An extremely supergaussian variable is one which is zero most of the time, only rarely obtaining other values. Thus one could use an 'improper' log-density defined as follows: $f(u) = 0$ for $u = 0$ and $f(u) = M$ for $u \neq 0$, where M is a very large constant. Using such a log-density, at most one of the $s_i(t)$ is non-zero, for practically every t. Note that the case where all the $s_i(t)$ are zero corresponds to the case where only Gaussian noise is observed. Since the estimation of $\mathbf{A}$ is not strongly affected by occasionally inputting Gaussian noise as $\mathbf{x}$, we may simplify the situation even further by assuming that for any given $\mathbf{x}(t)$, *exactly one of the s_i is non-*

zero. This means that we are using a competitive winner-take-all mechanism in the representation of the $\mathbf{x}(t)$. The second term in the likelihood is then constant, and one only needs to minimize the reconstruction error (i.e., the first term) subject to the constraint that only one of the $s_i(t)$ is non-zero for given t, i.e., only one neuron is active at a given point of time.

Given $\hat{\mathbf{A}}$, one can thus determine the $\hat{\mathbf{s}}(t)$ using the competitive mechanism. Then the algorithm in the preceding section leads to the following iteration for updating the columns $\hat{\mathbf{a}}_1, ..., \hat{\mathbf{a}}_n$ of $\hat{\mathbf{A}}$:

(i) For each vector $\hat{\mathbf{a}}_i(k)$, collect the set S_i of all those training samples $\mathbf{x}(t)$ such that $|\hat{\mathbf{a}}_i(k)^T\mathbf{x}(t)| \geq |\hat{\mathbf{a}}_j(k)^T\mathbf{x}(t)|$ for all $j \neq i$.
(ii) Let $\hat{\mathbf{a}}_i^*(k) = \sum_{\mathbf{x}(t)\in S_i} \mathbf{x}(t)\,(\hat{\mathbf{a}}_i(k)^T\mathbf{x}(t))$ and $\hat{\mathbf{a}}_i(k+1) = \hat{\mathbf{a}}_i^*(k)/\|\hat{\mathbf{a}}_i^*(k)\|$

where k is the iteration index. This algorithm is basically a modification of the classical k-means algorithm. The main difference is that instead of vectors, we are 'quantizing' 1-D subspaces. Note that since the improper density does not determine the variance of the s_i, we normalize instead the vectors $\hat{\mathbf{a}}_i(k)$. Note also that the estimates of the $\mathbf{s}(t)$ obtained by winner-take-all learning may not be very good: the $\mathbf{s}(t)$ should be estimated separately after the estimation of $\mathbf{A}$, using the methods in Section 3. To improve the convergence of this learning rule, it may be advisable to sphere (or whiten) the data $\mathbf{x}$.

An important feature of this competitive learning rule is that it was not assumed in the derivation that the number of observed mixtures x_i be at least as large as the number of the independent components. Thus it allows the estimation of $\mathbf{A}$ even in the case where there are more independent components than mixtures, i.e., $n > m$.

5.3 An anti-competitive mechanism for subgaussian components

For estimating $\mathbf{A}$ in the case of *subgaussian* independent components, we use an extremely subgaussian 'density', which corresponds to the binary distribution: $s_i = \pm 1$ with equal probabilities. Thus we constrain the s_i to have one of the two values: -1 and $+1$. Again, the second term in the likelihood is constant, and one only needs to minimize the reconstruction error (i.e., the first term) subject to $s_i(t) = \pm 1$. To simplify the situation, assume that the data is sphered and that the matrix $\hat{\mathbf{A}}$ is constrained to be orthogonal (which implies that $n \leq m$). Then it is obvious that the optimal reconstruction is obtained when $\hat{\mathbf{s}}(t) = \text{sign}(\mathbf{A}^T\mathbf{x}(t))$, where the sign operator is applied separately on each component of its argument. Then the step 3 of the algorithm in Section 4 has the form

$$\hat{\mathbf{A}}_k = E\{\mathbf{x}(\text{sign}(\mathbf{x}^T\hat{\mathbf{A}}_{k-1})\} \tag{8}$$

This learning rule might be called anti-competitive, as every unit participates in the representation of $\mathbf{s}(t)$ with the same amount.

6 Application for denoising

The framework given in Section 3 has applications beyond the simple estimation problem considered here.

Assume that we estimate the noise-free independent components by the methods introduced in Section 3, obtaining the estimates $\hat{\mathbf{s}}(t)$. We could reconstruct the original data by

$$\hat{\mathbf{x}} = \hat{\mathbf{A}}\hat{\mathbf{s}}. \tag{9}$$

It is then reasonable to assume that $\hat{\mathbf{x}}$ contains less noise than the original $\mathbf{x}$. This gives a new approach to denoising of nongaussian random vectors. The resulting method is called sparse code shrinkage and is described in detail in [10,9]. In fact, it is shown in [9] that it is not really necessary that the data follows the ICA model: The crucial assumption is that the data is nongaussian, which seems to be true for many real-world signals. Thus the method is of considerable generality.

7 Simulation results

We applied the methods of our paper for blind separation of 3 i.i.d. source signals (or independent components) from 3 noisy mixtures. The source signals are shown in Fig. 2, and the noisy mixtures in Fig. 3. The competitive learning rule in Section 5.2 was used for estimating $\mathbf{A}$. Approximately 10 iterations were required for convergence. The linearly separated signals $\hat{\mathbf{A}}^{-1}\mathbf{x}(t)$ are shown in Fig. 4. The errors in these linearly separated signals are *not* due to errors in the estimation of $\mathbf{A}$, since $\mathbf{A}$ was estimated correctly up to 4 decimal places. Instead, the errors in Fig. 4 are due to noise and linear reconstruction. Applying the shrinkage operator on the linear reconstruction, one obtained an approximation of the optimal nonlinear reconstructions of the source signals, depicted in Fig. 5. Clearly, this non-linear reconstruction gives better estimates of the original source signals in the presence of noise.

Next, we performed estimation of $\mathbf{A}$ in the case where the number of independent components was larger than the number of observed mixtures. The independent components had supergaussian distributions, and thus the competitive learning in Section 5.2 was used again. To validate the results the

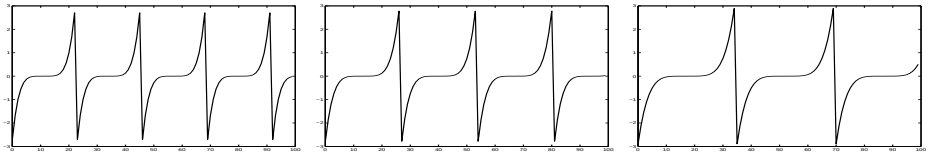

Fig. 2. Original supergaussian source signals.

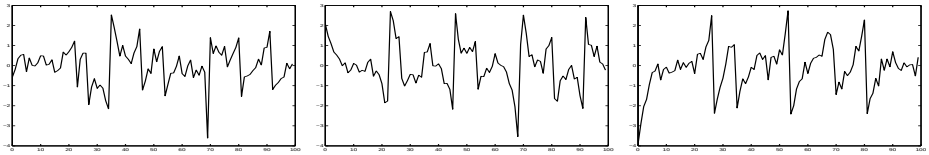

Fig. 3. Mixtures of supergaussian source signals, with noise added.

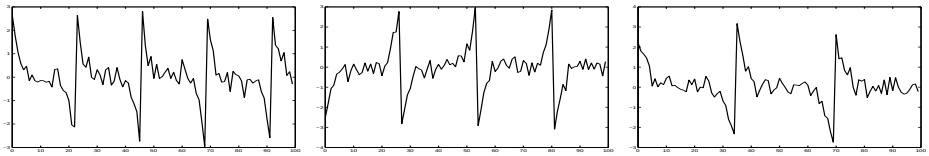

Fig. 4. Linearly separated supergaussian source signals.

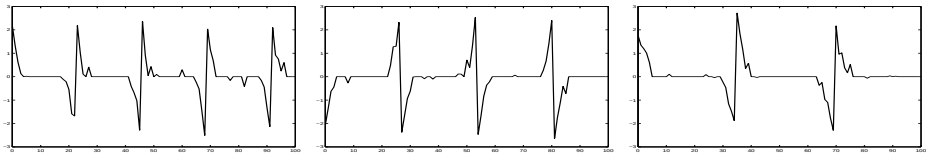

Fig. 5. Non-linear reconstructions of the supergaussian source signals.

matrix $\mathbf{A}^T\hat{\mathbf{A}}$ was computed after convergence. The matrices were here normalized so that each column had unit norm. This gave the following:

$$\mathbf{A}^T\hat{\mathbf{A}} = \begin{bmatrix} -0.1001 & -0.0760 & -0.9999 & -0.0642 \\ -0.9998 & 0.6031 & -0.0970 & 0.5187 \\ 0.6032 & -0.9998 & -0.0544 & 0.3567 \\ 0.4887 & 0.3586 & -0.0522 & -0.9998 \end{bmatrix} \tag{10}$$

Every column has exactly one entry whose absolute value is practically equal to one. This shows that the columns of $\hat{\mathbf{A}}$ converged to the directions of the columns of $\mathbf{A}$, which shows that the algorithm converged properly.

Finally, we performed simulations to validate the algorithm in (8). The results are given in Figs. 6–9. The original distributions of the independent components were uniform or binary. The corresponding nonlinear reconstructions correspond to the truncation operator (as given above) and the 'sign' nonlinearity. Clearly, the method was able to estimate the mixing matrix, and reduce noise from the linear estimates of the source signals.

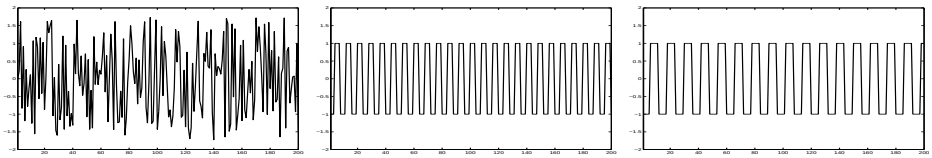

Fig. 6. Original subgaussian source signals.

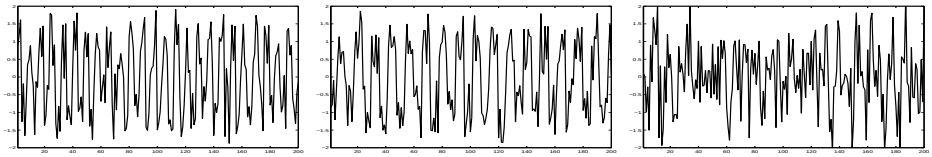

Fig. 7. Mixtures of subgaussian source signals, with noise added.

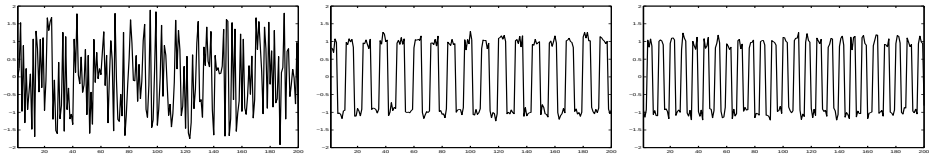

Fig. 8. Linearly separated subgaussian source signals.

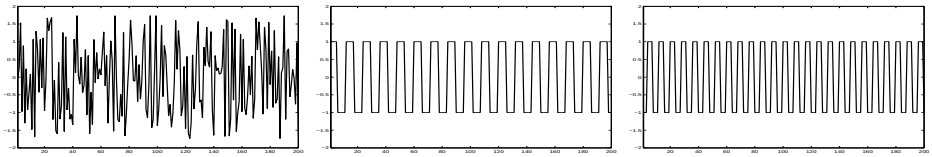

Fig. 9. Non-linear reconstructions of the subgaussian source signals.

8 Discussion

A maximum likelihood approach was taken to the estimation of the ICA data model. It was shown how to estimate jointly the mixing matrix and the independent components. This leads, in general, to a very complex minimization problem. Therefore, simpler approximations were introduced. It was shown how the presence of noise leads to a non-linear relationship between the observed variables and the estimates of the independent components, and methods for approximating this non-linear relation were given. A simple alternating variables method was then proposed using these approximations.

To simplify the estimation even further, it was shown how the estimation can be performed by competitive and anti-competitive learning rules for super- and subgaussian data, respectively. This shows an interesting connection between competitive learning and ICA.

Little work on maximum likelihood estimation of ICA in the presence of noise has been done previously. In [3] the problem was treated for discrete-valued sources. A method more closely related to ours was introduced in [17], in which an EM algorithm was proposed for estimation of the noisy ICA model. Our

methods are considerably simpler and computationally less demanding than the one proposed in [17]. In particular, the complexity of the EM algorithm is exponential as a function of the number of independent components; our algorithm has polynomial complexity. The price to pay is the loss of generality. The approximations introduced in Sections 3 and 5 may not work in all circumstances. Moreover, in this paper, the covariance matrix of the noise was assumed to be known, whereas in [17], it was estimated as part of the algorithm. Future work may reveal simple ways of estimating the covariance matrix in our framework. Other recent work on the problem of noisy ICA is reported by Cichocki et al [5] in this issue.

A Derivations

Taking the negative of the gradient of (2) with respect to $\mathbf{s}(t)$, we obtain

$$\mathbf{A}^T\Sigma^{-1}\mathbf{A}\mathbf{s}(t) - \mathbf{A}^T\Sigma^{-1}\mathbf{x}(t) + f'(\mathbf{s}(t)). \tag{A.1}$$

The constraints of unit variance are equivalent to constraining the norms of the vectors $\mathbf{S}_i = (s_i(1), s_i(2), ..., s_i(T))$ to equal one. This corresponds to the Lagrangian term

$$\sum_i \lambda_i(\|\mathbf{S}_i\|^2 - 1) \tag{A.2}$$

which gives the gradient of the Lagrangian of (2) with respect to the whole data $\mathbf{s}(1), ..., s(T)$; this can be written separately for each t as

$$\mathbf{A}^T\Sigma^{-1}\mathbf{A}\mathbf{s}(t) - \mathbf{A}^T\Sigma^{-1}\mathbf{x}(t) + f'(\mathbf{s}(t)) - \Lambda\mathbf{s}(t). \tag{A.3}$$

which implies (3). To derive (5), note that the assumption $\Sigma = \sigma^2\mathbf{A}\mathbf{A}^T$ implies

$$\frac{1}{\sigma^2}\hat{\mathbf{s}}(t) - \frac{1}{\sigma^2}\hat{\mathbf{A}}^{-1}\mathbf{x}(t) + f'(\hat{\mathbf{s}}(t)) - \hat{\mathbf{s}}(t) = 0. \tag{A.4}$$

which gives (6). Finally, to derive (7), it is enough to plot the function in (6), with $f'(u) = \sqrt{2}\text{sign}(u)$. The inverse can be obtained by reflecting the graph with respect to the axis $x = y$.

References

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

[2] A.J. Bell and T.J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37:3327–3338, 1997.

[3] A. Belouchrani and J.-F. Cardoso. Maximum likelihood source separation by the expectation-maximization technique: deterministic and stochastic implementation. In *Proc. NOLTA*, pages 49–53, 1995.

[4] J.-F. Cardoso and B. Hvam Laheld. Equivariant adaptive source separation. *IEEE Trans. on Signal Processing*, 44(12):3017–3030, 1996.

[5] A. Cichocki, S. C. Douglas, and S. i. Amari. Robust techniques for independent component anlaysis with noisy data. *Neurocomputing*, 1998. In this issue.

[6] A. Cichocki and R. Unbehauen. *Neural Networks for Signal Processing and Optimization*. Wiley, 1994.

[7] P. Comon. Independent component analysis – a new concept? *Signal Processing*, 36:287–314, 1994.

[8] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society ser. B*, 57:301–337, 1995.

[9] A. Hyvärinen. Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation. Technical report, Helsinki University of Technology, Laboratory of Computer and Information Science, 1998. To appear.

[10] A. Hyvärinen, P. Hoyer, and E. Oja. Sparse code shrinkage for image denoising. In *Proc. IEEE Int. Joint Conf. on Neural Networks*, Anchorage, Alaska, 1998. To appear.

[11] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[12] A. Hyvärinen and E. Oja. Independent component analysis by general nonlinear Hebbian-like learning rules. *Signal Processing*, 64(3):301–313, 1998.

[13] C. Jutten and J. Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[14] J. Karhunen, A. Hyvärinen, R. Vigario, J. Hurri, and E. Oja. Applications of neural blind separation to signal and image processing. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'97)*, pages 131–134, Munich, Germany, 1997.

[15] J. Karhunen, E. Oja, L. Wang, R. Vigario, and J. Joutsensalo. A class of neural networks for independent component analysis. *IEEE Trans. on Neural Networks*, 8(3):486–504, 1997.

[16] J. Karhunen and P. Pajunen. Blind source separation using least-squares type adaptive algorithms. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'97)*, pages 3048–3051, Munich, Germany, 1997.

[17] E. Moulines, J.-F. Cardoso, and E. Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'97)*, pages 3617–3620, Munich, Germany, 1997.

[18] E. Oja. The nonlinear PCA learning rule in independent component analysis. *Neurocomputing*, 17(1):25–46, 1997.

[19] B. A. Olshausen and D. J. Field. Natural image statistics and efficient coding. *Network*, 7(2):333–340, May 1996.

[20] J. J. K. O'Ruanaidh and W. J. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing*. Springer, 1996.

[21] D.-T. Pham, P. Garrat, and C. Jutten. Separation of a mixture of independent sources through a maximum likelihood approach. In *Proc. EUSIPCO*, pages 771–774, 1992.