BLIND SEPARATION OF POSITIVE SOURCES USING NON-NEGATIVE PCA

Erkki Oja

Neural Networks Research Centre
Helsinki University of Technology
P.O.Box 5400, 02015 HUT, Finland
erkki.oja@hut.fi

Mark Plumbley

Department of Electrical Engineering
Queen Mary, University of London
Mile End Road
London E1 4NS, United Kingdom
mark.plumbley@elec.qmul.ac.uk

ABSTRACT

The instantaneous noise-free linear mixing model in independent component analysis is largely a solved problem under the usual assumption of independent nongaussian sources and full rank mixing matrix. However, with some prior information on the sources, like positivity, new analysis and perhaps simplified solution methods may yet become possible. In this paper, we consider the task of independent component analysis when the independent sources are known to be non-negative and well-grounded, which means that they have a non-zero pdf in the region of zero. We propose the use of a 'Non-Negative PCA' algorithm which is a special case of the nonlinear PCA algorithm, but with a rectification nonlinearity, and we show that this algorithm will find such non-negative well-grounded independent sources. Although the algorithm has proved difficult to analyze in the general case, we give an analytical convergence result here, complemented by a numerical simulation which illustrates its operation.

1. INTRODUCTION

The problem of independent component analysis (ICA) has been studied by many authors in recent years (for a review, see e.g. [1]). In the simplest form of ICA we assume that we have a sequence of observations $\{\mathbf{x}(k)\}$ which are samples of a random observation vector $\mathbf{x}$ generated according to

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

where $\mathbf{s} = (s_1, \ldots, s_n)^T$ is a vector of real independent random variables (the *sources*), all but perhaps one of them nongaussian, and $\mathbf{A}$ is a nonsingular $n \times n$ real *mixing matrix*. The task in ICA is to identify $\mathbf{A}$ given just the observation sequence, using the assumption of independence of the s_is, and hence to construct an unmixing matrix $\mathbf{B} = \mathbf{R}\mathbf{A}^{-1}$ giving $\mathbf{y} = \mathbf{B}\mathbf{x} = \mathbf{B}\mathbf{A}\mathbf{s} = \mathbf{R}\mathbf{s}$ where $\mathbf{R}$ is a matrix which permutes and scales the sources. Typically we assume that the sources have unit variance, with any scaling factor being absorbed into the mixing matrix $\mathbf{A}$, so $\mathbf{y}$ will be a permutation of the $\mathbf{s}$ with just a sign ambiguity.

Common cost functions for ICA are based on maximizing nongaussianities of the elements of $\mathbf{y}$ and they may involve higher-order cumulants such as kurtosis. The observations $\mathbf{x}$ are often assumed to be zero-mean, or transformed to be so, and are commonly pre-whitened by some matrix $\mathbf{z} = \mathbf{V}\mathbf{x}$ so that $E\{\mathbf{z}\mathbf{z}^T\} = \mathbf{I}$ before an optimization algorithm is applied to find the separating matrix.

Recently, one of the current authors considered an additional assumption on the sources: that they are *non-negative* as well as independent [2, 3]. Non-negativity is a natural condition for many real-world applications, for example in the analysis of images [4, 5], text [6], or air quality [7]. The constraint of non-negative sources, perhaps with an additional constraint of non-negativity on the mixing matrix $\mathbf{A}$, is often known as *positive matrix factorization* [8] or *non-negative matrix factorization* [9]. We refer to the combination of non-negativity and independence assumptions on the sources as *non-negative independent component analysis*.

Non-negativity of sources can provide us with an alternative way of approaching the ICA problem, as follows. We call a source s_i *non-negative* if $\Pr(s_i < 0) = 0$, and such a source will be called *well-grounded* if $\Pr(s_i < \delta) > 0$ for any $\delta > 0$, i.e. that s_i has non-zero pdf all the way down to zero. One of the authors proved the following [2]:

Theorem 1. Suppose that $\mathbf{s}$ is a vector of non-negative well-grounded independent unit-variance sources s_i, $i = 1, ..., n$, and $\mathbf{y} = \mathbf{U}\mathbf{s}$ where $\mathbf{U}$ is a square orthonormal rotation, i.e. $\mathbf{U}^T\mathbf{U} = \mathbf{I}$. Then $\mathbf{U}$ is a permutation matrix, i.e. the elements y_j of $\mathbf{y}$ are a permutation of the sources s_i, if and only if all y_j are non-negative.

This result can be used for a simple solution of the non-negative ICA problem: note that $\mathbf{y} = \mathbf{U}\mathbf{s}$ can also be written as $\mathbf{y} = \mathbf{W}\mathbf{z}$ with $\mathbf{z}$ the pre-whitened observation vector and $\mathbf{W}$an unknown orthogonal (rotation) matrix. It therefore

suffices to *find an orthogonal matrix* $\mathbf{W}$ *for which* $\mathbf{y} = \mathbf{W}\mathbf{z}$ *is non-negative*. This brings the additional benefit over other ICA methods that we know of that, if successful, we always have a positive permutation of the sources, since both the $\mathbf{s}$ and $\mathbf{y}$ are non-negative. The sign ambiguity present in usual ICA vanishes here.

In the next Section 2, we present the whitening for non-zero mean observations and further illustrate by a simple example why a rotation into positive outputs y_i will give the sources. Then, in Section 3 we show the relation of non-negativity to "non-negative PCA" and in Section 4 give a gradient algorithm, whose global convergence is proven. Section 5 relates this orthogonalized algorithm to the non-orthogonal "nonlinear PCA" learning rule previously introduced by one of the authors. Section 6 illustrates the non-negative PCA principle by a pictorial example, and Section 7 gives some conclusions.

2. PRE-WHITENING AND AXIS ROTATIONS

In order to reduce the ICA problem to one of finding the correct orthogonal rotation, the first stage in our ICA process is to *whiten* the observed data $\mathbf{x}$. This gives

$$\mathbf{z} = \mathbf{V}\mathbf{x} \tag{2}$$

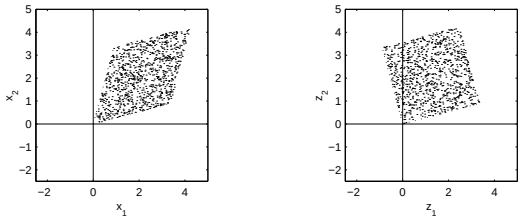
where the $n \times n$ real whitening matrix $\mathbf{V}$ is chosen so that $\mathbf{\Sigma_z} = E\{(\mathbf{z} - \bar{\mathbf{z}})(\mathbf{z} - \bar{\mathbf{z}})^T\} = \mathbf{I}_n$, with $\bar{\mathbf{z}} = E\{\mathbf{z}\}$. If $\mathbf{E}$ is the orthogonal matrix of eigenvectors of $\mathbf{\Sigma_x} = E\{(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T\}$ and $\mathbf{D} = \text{diag}(d_1, \ldots, d_n)$ is the diagonal matrix of corresponding eigenvalues, so that $\mathbf{\Sigma_x} = \mathbf{E}\mathbf{D}\mathbf{E}^T$ and $\mathbf{E}^T\mathbf{E} = \mathbf{E}\mathbf{E}^T = \mathbf{I}_n$, then a suitable whitening matrix is $\mathbf{V} = \mathbf{\Sigma_x}^{-1/2} = \mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^T$ where

$$\mathbf{D}^{-1/2} = \text{diag}(d_1^{-1/2}, \ldots, d_n^{-1/2}).$$

$\mathbf{\Sigma_x}$ is normally estimated from the sample covariance [1]. Note that for non-negative ICA, we do not remove the mean of the data, since this would lose information about the non-negativity of the sources [2].

Suppose that our sources s_j have unit variance, such that $\mathbf{\Sigma_s} = \mathbf{I}_n$, and let $\mathbf{U} = \mathbf{V}\mathbf{A}$ be the $\mathbf{s}$-to-$\mathbf{z}$ transform. Then $\mathbf{I}_n = \mathbf{\Sigma_z} = \mathbf{U}\mathbf{\Sigma_s}\mathbf{U}^T = \mathbf{U}\mathbf{U}^T$ so $\mathbf{U}$ is an orthonormal matrix. It is therefore sufficient to search for a further orthonormal matrix $\mathbf{W}$ such that $\mathbf{y} = \mathbf{W}\mathbf{z} = \mathbf{W}\mathbf{U}\mathbf{s}$ is a permutation of the original sources $\mathbf{s}$.

Figure 1 illustrates the process of whitening for non-negative data in 2 dimensions. Whitening has succeeded in making the axes of the original sources orthogonal to each other (Fig. 1(b)), but there is a remaining orthonormal rotation ambiguity. A typical ICA algorithm might search for a rotation that makes the resulting outputs as non-gaussian as possible, for example by finding an extremum of kurtosis, since any sum of independent random variables will make the result 'more gaussian' [1].



(a) (b)

Fig. 1. Original data (a) is whitened (b) to remove 2nd order correlations.

However, Figure 1 immediately suggests another approach: we should search for a rotation where all of the data fits into the positive quadrant. As long as the distribution of the original sources is 'tight' down to the axes, then it is intuitively clear that this will be a unique solution, apart from a permutation and scaling of the axes. This explains why Theorem 1 works.

3. NON-NEGATIVITY AND NONLINEAR PCA

The result of Theorem 1 has a connection with Principal Component Analysis (PCA) as follows. Recall the classical result saying that, given an n dimensional vector $\mathbf{x}$, its k dimensional principal component subspace can be found by minimizing the representation error:

$$e_{MSE} = E\|\mathbf{x} - \mathbf{W}^T\mathbf{W}\mathbf{x}\|^2$$

where $\mathbf{W}$ is a $k \times n$ matrix. The minimum of e_{MSE} is given by a matrix with orthonormal rows, and the matrix $\mathbf{W}^T\mathbf{W}$ is then the projector on the dominant eigenvector subspace of $\mathbf{x}$, spanned by the k dominant eigenvectors of the covariance matrix of $\mathbf{x}$ [10].

If $k = n$, the criterion is meaningless, because the whole space is the principal component subspace and $\mathbf{W}^T\mathbf{W} = \mathbf{I}$. The error e_{MSE} attains the value zero. However, if $\mathbf{W}\mathbf{x}$ is replaced by a nonlinear function $g(\mathbf{W}\mathbf{x})$, applied element by element, then the problem changes totally: the representation error usually does not attain the value of zero any more even for $k = n$, and in some cases the minimization leads to independent components [11]. Let us write this nonlinear MSE criterion, first introduced by Xu [12], for the whitened vector $\mathbf{z}$ instead:

$$e_{MSE} = E\|\mathbf{z} - \mathbf{W}^T g(\mathbf{W}\mathbf{z})\|^2. \tag{3}$$

Let us choose for the nonlinearity g the *rectification nonlinearity*

$$g(y_i) = g_+(y_i) = \max(0, y_i)$$

which is zero for negative y_i, and y_i otherwise. Then the criterion (3) can be called "Non-Negative PCA" because for

the positive elements of $\mathbf{Wz}$ it coincides with usual PCA and for the negative elements g is zero, having no effect.

We are now ready to state

Theorem 2. Assume the n-element random vector $\mathbf{z}$ is a whitened linear mixture of non-negative well-grounded independent unit variance sources $s_i, ..., s_n$, and $\mathbf{y} = \mathbf{Wz}$ with $\mathbf{W}$ constrained to be a square orthogonal matrix. If $\mathbf{W}$ is obtained as the minimum of $E\|\mathbf{z} - \mathbf{W}^T g_+(\mathbf{Wz})\|^2$, then the elements of $\mathbf{y}$ will be a permutation of the original sources s_i.

Proof. Because $\mathbf{W}$ is square orthogonal, we get

$$\begin{aligned} e_{MSE} &= E\|\mathbf{z} - \mathbf{W}^T g_+(\mathbf{Wz})\|^2 && (4)\\ &= E\|\mathbf{Wz} - \mathbf{WW}^T g_+(\mathbf{Wz})\|^2 && (5)\\ &= E\|\mathbf{y} - g_+(\mathbf{y})\|^2 && (6)\\ &= \sum_{i=1}^{n} E\{[y_i - g_+(y_i)]^2\} && (7)\\ &= \sum_{i=1}^{n} E\{\min(0, y_i)^2\} && (8)\\ &= \sum_{i=1}^{n} E\{y_i^2 | y_i < 0\} P(y_i < 0). && (9) \end{aligned}$$

This is always non-negative and becomes zero if and only if each y_i is non-negative with probability one.

On the other hand, because $\mathbf{y} = \mathbf{Wz}$ with $\mathbf{W}$ orthogonal, then it also holds that $\mathbf{y} = \mathbf{Us}$ with $\mathbf{U}$ orthogonal. Theorem 1 now implies that $\mathbf{y}$ must be a permutation of $\mathbf{s}$ **QED**.

4. A CONVERGING GRADIENT ALGORITHM

Theorem 2 leads us naturally to consider the use a gradient algorithm for minimizing (3) under the orthogonality constraint. This problem was considered by one of the authors [13], and the ensuing gradient descent algorithm becomes

$$\Delta\mathbf{W} = -\eta[f(\mathbf{y})\mathbf{y}^T - \mathbf{y}f(\mathbf{y})^T]\mathbf{W} \qquad (10)$$

where now

$$f(y_i) = \min(0, y_i) \qquad (11)$$

and η is the positive learning rate.

The skew-symmetric form of the matrix $f(\mathbf{y})\mathbf{y}^T - \mathbf{y}f(\mathbf{y})^T$ ensures that $\mathbf{W}$ tends to stay orthogonal from step to step, although to fully guarantee orthogonality, an explicit orthonormalization of the rows of $\mathbf{W}$ should be done from time to time.

Instead of analyzing the learning rule directly, let us look at the averaged differential equation corresponding to the discrete-time stochastic algorithm (10). It becomes

$$\frac{d\mathbf{W}}{dt} = -\mathbf{MW} \qquad (12)$$

where we have denoted the continuous-time deterministic solution also by $\mathbf{W}$, and the elements μ_{ij} of matrix $\mathbf{M}$ are

$$\mu_{ij} = E\{\min(0, y_i)y_j - y_i \min(0, y_j)\}. \qquad (13)$$

Note that $\mathbf{M}$ is a nonlinear function of the solution $\mathbf{W}$, because $\mathbf{y} = \mathbf{Wz}$. Yet, we can formally write the solution of (12) as

$$\mathbf{W}(t) = \exp[-\int_0^t \mathbf{M}(s)ds]\mathbf{W}(0). \qquad (14)$$

The solution $\mathbf{W}(t)$ is always an orthogonal matrix, if $\mathbf{W}(0)$ is orthogonal. This can be shown as follows:

$$\begin{aligned} &\mathbf{W}(t)\mathbf{W}(t)^T = \\ &\exp[-\int_0^t \mathbf{M}(s)ds]\mathbf{W}(0)\mathbf{W}(0)^T \exp[-\int_0^t \mathbf{M}(s)^T ds] \\ = \;& \exp[-\int_0^t (\mathbf{M}(s) + \mathbf{M}(s)^T)ds]. \end{aligned}$$

But matrix $\mathbf{M}$ is skew-symmetric, hence $\mathbf{M}(s) + \mathbf{M}(s)^T = 0$ for all s and $\mathbf{W}(t)\mathbf{W}(t)^T = \exp[0] = \mathbf{I}$.

We can now analyze the stationary points of (12) and their stability in the class of orthogonal matrices. The stationary points (for which $\frac{d\mathbf{W}}{dt} = 0$) are easily solved. They must be the roots of the equation $\mathbf{M} = 0$. We see that if all y_i are positive or all of them are negative, then $\mathbf{M} = 0$. Namely, if y_i and y_j are both positive, then $\min(0, y_i)$ and $\min(0, y_j)$ in (13) are both zero. If they are both negative, then $\min(0, y_i) = y_i$ and $\min(0, y_j) = y_j$ and the two terms in (13) cancel out. Thus, in these two cases $\mathbf{W}$ is a stationary point. The case when all y_i are positive corresponds to the minimum value (zero) of the cost function e_{MSE}. By Theorem 1, $\mathbf{y}$ is then a permutation of $\mathbf{s}$, which is the correct solution we are looking for. We would hope that this stationary point would be the only stable one, because then the ode will converge to it.

The case when all the y_i are negative corresponds to the maximum value of e_{MSE}, equal to $\sum_{i=1}^{n} E\{y_i^2\} = n$. As it is stationary, too, we have to consider the case when it is taken as the initial value in the ode.

In all other cases, at least some of the y_i have opposite signs. Then $\mathbf{M}$ is not zero and $\mathbf{W}$ is not stationary, as seen from (14).

We could look at the local stability of the two stationary points. However, we can do even better and perform a global analysis. It turns out that (4) is in fact a Lyapunov function for the matrix flow (12); it is strictly decreasing always when $\mathbf{W}$ changes according to the ode (12), except at the stationary points. Let us prove this in the following.

Theorem 3. If $\mathbf{W}$ follows the ode (12), then $\frac{de_{MSE}}{dt} < 0$, except at the point when all y_i are non-negative or all are non-positive.

Proof. Consider the ith term in the sum e_{MSE}. Denoting it by e_i, we have $e_i = E\{(\min(0, y_i))^2\}$ whose derivative with respect to y_i is easily shown to be $2E\{\min(0, y_i)\}$. If $\mathbf{w}_i^T$ is the ith row of matrix $\mathbf{W}$, then $y_i = \mathbf{w}_i^T\mathbf{z}$. Thus

$$\frac{de_i}{dt} = \frac{de_i}{dy_i}\frac{dy_i}{dt} = 2E\{\min(0, y_i)(\frac{d\mathbf{w}_i^T}{dt}\mathbf{z})\}. \quad (15)$$

From the ode (12) we get

$$\frac{d\mathbf{w}_i^T}{dt} = -\sum_{k=1}^{n} \mu_{ik}\mathbf{w}_k^T$$

with μ_{ik} given in (13). Substituting now this in (15) gives

$$\begin{aligned}\frac{de_i}{dt} &= -2\sum_{k=1}^{n} \mu_{ik}E\{\min(0, y_i)y_k\} \\ &= -2\sum_{k=1}^{n} E^2\{\min(0, y_i)y_k\} \\ &+ 2\sum_{k=1}^{n} E\{\min(0, y_k)y_i\}E\{\min(0, y_i)y_k\}.\end{aligned}$$

If we denote $\alpha_{ik} = E\{\min(0, y_i)y_k\}$, we have

$$\frac{de_{MSE}}{dt} = \sum_{i=1}^{n}\frac{de_i}{dt} = 2[-\sum_{i=1}^{n}\sum_{k=1}^{n}\alpha_{ik}^2 + \sum_{i=1}^{n}\sum_{k=1}^{n}\alpha_{ik}\alpha_{ki}].$$

By the Cauchy-Schwartz inequality, this is strictly negative unless $\alpha_{ik} = \alpha_{ki}$, and thus e_{MSE} is decreasing.

We still have to look at the condition that $\alpha_{ik} = \alpha_{ki}$ and show that this implies non-negativity or non-positivity for all the y_i.

Now, because $\mathbf{y} = \mathbf{U}\mathbf{s}$ with $\mathbf{U}$ orthogonal, each y_i is a projection of the positive source vector $\mathbf{s}$ on one of n orthonormal rows $\mathbf{u}_i^T$ of $\mathbf{U}$. If the vectors $\mathbf{u}_i$ are aligned with the original coordinate axes, then the projections of $\mathbf{s}$ on them are non-negative. For any rotation that is not aligned with the coordinate axes, one of the vectors $\mathbf{u}_i$ (or $-\mathbf{u}_i$) must be in the positive octant, due to the orthonormality of the vectors. Without loss of generality, assume that this vector is $\mathbf{u}_1$; then it holds that $P(y_1 = \mathbf{u}_1^T\mathbf{s} \geq 0) = 1$ (or 0). But if $P(y_1 \geq 0) = 1$, then $\min(0, y_1) = 0$ and $\alpha_{1k} = E\{\min(0, y_1)y_k\} = 0$ for all k. If symmetry holds for the α_{ij}, then also $\alpha_{k1} = E\{\min(0, y_k)y_1\} = E\{y_1y_k|y_k \leq 0\}P(y_k \leq 0) = 0$. But y_1 is non-negative, so $P(y_k \leq 0)$ must be zero, too, for all k. The same argument carries over to the case when $P(y_1 \geq 0) = 0$, which implies that if one y_i is non-negative, then all y_k must be non-negative in the case of symmetrical α_{ij} **QED**.

The behaviour of the learning rule (10) is now well understood. Even if the starting point would happen to be the "bad" stationary point in which all y_i are non-positive, then numerical errors will deviate the solution from this point and the cost function e_{MSE} starts to decrease. This was proven only for the continuous-time averaged version of the learning rule; the exact connection between this and the discrete-time on-line algorithm has been clarified in the theory of stochastic approximation. See e.g. [14]. The cost function is decreasing and non-negative and will convergence to the stationary minimum, corresponding to all non-negative y_i. By Theorem 1, these must be a permutation of the original sources s_j which therefore have been found.

5. RELATION TO THE NONLINEAR PCA LEARNING RULE

For the general nonlinear MSE criterion, given in (3), the full matrix gradient is [15]

$$\frac{\partial e_{MSE}}{\partial \mathbf{W}} = -E\{\mathbf{F}(\mathbf{y})\mathbf{W}\mathbf{r}\mathbf{z}^T + g(\mathbf{y})\mathbf{r}^T\} \quad (16)$$

where $\mathbf{r}$ is the representation error

$$\mathbf{r} = \mathbf{z} - \mathbf{W}^T g(\mathbf{W}\mathbf{z})$$

and

$$\mathbf{F}(\mathbf{y}) = \text{diag}(g'(y_1), ..., g'(y_n)).$$

Now, consider the present case of $g(y) = g_+(y)$ and look at the first term on the right hand side of (16). We have

$$\begin{aligned}\mathbf{F}(\mathbf{y})\mathbf{W}\mathbf{r}\mathbf{z}^T &= \mathbf{F}(\mathbf{y})\mathbf{W}[\mathbf{z} - \mathbf{W}^T g_+(\mathbf{W}\mathbf{z})]\mathbf{z}^T \\ &= \mathbf{F}(\mathbf{y})[\mathbf{y} - g_+(\mathbf{y})]\mathbf{z}^T.\end{aligned}$$

The ith element of vector $\mathbf{F}(\mathbf{y})[\mathbf{y} - g(\mathbf{y})]$ is $g'_+(y_i)[y_i - g_+(y_i)]$ which is clearly zero: if $y_i \leq 0$, then $g'_+(y_i) = 0$, and if $y_i > 0$, then this term becomes $1 \times [y_i - y_i] = 0$.

This means that the first term in the gradient vanishes altogether and what remains is the term

$$-g_+(\mathbf{y})\mathbf{r}^T = -g_+(\mathbf{y})(\mathbf{z} - \mathbf{W}^T g_+(\mathbf{y}))^T.$$

This shows that the on-line gradient descent rule for the e_{MSE} criterion can also be written as

$$\Delta\mathbf{W} = \eta g_+(\mathbf{y})(\mathbf{z} - \mathbf{W}^T g_+(\mathbf{y}))^T. \quad (17)$$

This has the form of the Nonlinear PCA learning rule, earlier suggested by one of the authors in [11].

Note that in the analysis above we used the assumption that matrix $\mathbf{W}$ stays orthogonal. This is strictly not true in the gradient algorithm, unless an explicit orthogonalization is performed at each step. The approximation is the better, the closer $\mathbf{W}$ is to orthogonality. It can be shown that stationary points of the gradient algorithm, even without explicit orthogonalizations, will be orthogonal matrices; thus

asymptotically, the orthogonality assumption holds and the approximating gradient coincides with the exact gradient. For computational reasons it may therefore be easier to use algorithm (17) instead of (10) for finding the positive independent components.

6. EXPERIMENTS

We illustrate the operation of non-negative PCA using a blind image separation problem (see e.g. [16]). This is suitable for non-negative ICA, since the source images have non-negative pixel values.

The original images used in this section are shown in Fig. 2. They are square 128×128 images (downsampled by a factor of 4 from the original 512×512 images) with integer pixel intensities between 0 and 255 inclusive, which were then scaled to unit variance. Each source sequence $s_j(k),\ j = 1, 2, 3$ is considered to be the sequence of pixel values obtained as we scan across the image from top left ($k = 0$) to bottom right ($k = 128^2$).

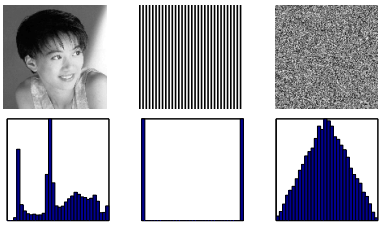


Fig. 2. Source images and histograms used for the non-negative ICA algorithms. *(The images were kindly supplied by Włodzimierz Kasprzak and Andrzej Cichocki.)*

We found that face images tended to have significant correlation with other face images, breaking the independence assumption of ICA methods. Consequently we used one face and two artificial images as the sources for these demonstrations.

Note that the histograms indicate that the face image has a non-zero minimum value, which does violate our assumption that the sources are *well-grounded* [2]. Nevertheless, we will see that we will get reasonable (although not perfect) separation performance from the nonnegative ICA algorithm.

To measure the separation performance of the algorithm, we use two performance measures: first, the nonnegative reconstruction error

$$e_{NNR} = \frac{1}{np}\|\mathbf{Z} - \mathbf{W}^T\mathbf{Y}_+\|_{\mathrm{F}}^2 \qquad (18)$$

where $\mathbf{Z}$ and $\mathbf{Y}_+$ are the matrices whose columns are the $p = 128^2$ pre-whitened observation vectors $\mathbf{z}$ of dimension $n = 3$ and the rectified versions of vectors $\mathbf{y} = \mathbf{W}\mathbf{z}$, respectively. Thus, this is the sample version of the error in (4). The other performance measure is the cross-talk error

$$e_{XT} = \frac{1}{n^2}\|\mathrm{abs}(\mathbf{WVA})^T\mathrm{abs}(\mathbf{WVA}) - \mathbf{I}_n\|_{\mathrm{F}}^2 \qquad (19)$$

where $\mathrm{abs}(\mathbf{WVA})$ is the matrix of absolute values of the elements of $\mathbf{WVA}$. This measure is zero only if $\mathbf{y} = \mathbf{WVAs}$ is a permutation of the sources, i.e. only if the sources have been successfully separated.

Fig. 3 gives an example learning curve for the non-negative PCA algorithm (17) of section 5. The learning rate was

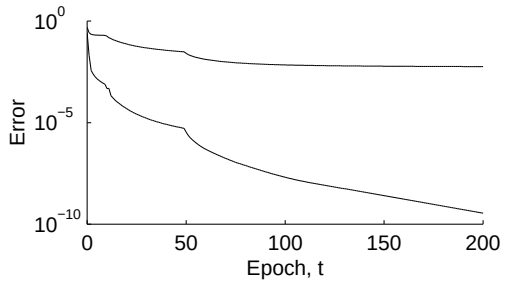


Fig. 3. Learning curve for the non-negative PCA algorithm, showing nonnegative reconstruction error (lower curve), and cross-talk error (upper curve).

manually adjusted to improve the convergence time, using $\eta = 10^4$ initially, $\eta = 10^3$ for $10 \le t < 50$, and $\eta = 10^2$ for $t \ge 50$. As the algorithm progresses, the images are successfully separated, and the histograms become less Gaussian and closer to those of the original sources (Fig. 4). As a non-negative algorithm, we never see inverted sources recovered, such as we might expect from traditional ICA algorithms.

From the curves, we see that the reconstruction error is decreasing steadily after the initial stage. However, the crosstalk error, measuring the distance away from separation, reaches a minimum of 5.73×10^{-3} after 200 epochs.

7. DISCUSSION

We have considered the problem of *Non-Negative ICA*, that is, independent component analysis where the sources are known to be non-negative. Elsewhere, one of us introduced algorithms to solve this based on the use of orthogonal rotations, related to Stiefel manifold approaches [3].

In this paper we considered gradient-based algorithms operating on pre-whitened data, related to the 'nonlinear PCA' algorithms investigated by one of the present authors [11, 13]. We refer to these algorithms, which use a rectification nonlinearity, as *Non-Negative PCA* algorithms. Theoretical analysis of algorithm (10), which includes a matrix of

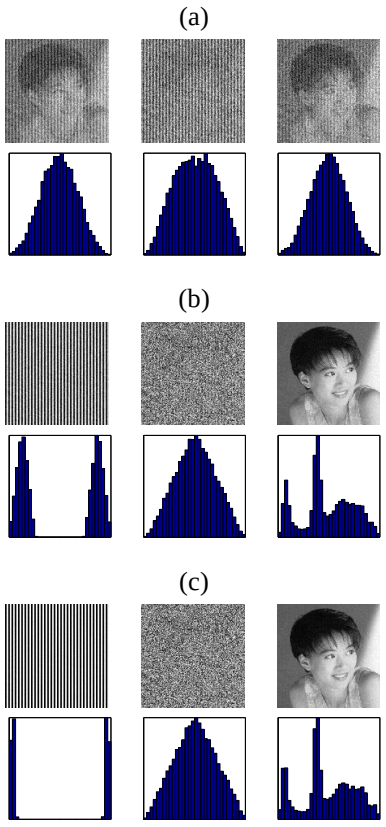


Fig. 4. Image separation process for the non-negative PCA algorithm, showing (a) the initial state, and progress after (b) 50 epochs and (c) 200 epochs.

skew-symmetric form, shows that it will tend to find a permutation of the non-negative sources. The related algorithm (17) is simpler in form, but is more difficult to analyze. Nevertheless, simulations indicate that this also reliably finds a permutation of the non-negative sources.

Acknowledgements. Andrzej Cichocki and Włodzimierz Kasprzak kindly supplied the images used in section 6. Part of this work was undertaken while the second author was visiting the Neural Networks Research Centre at the Helsinki University of Technology, supported by a Leverhulme Trust Study Abroad Fellowship, and this work is also supported by grant GR/R54620 from the UK Engineering and Physical Sciences Research Council as well as by the project New Information Processing Principles, 44886, of the Academy of Finland.

8. REFERENCES

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[2] M. D. Plumbley, "Conditions for nonnegative independent component analysis," *IEEE Signal Processing Letters*, vol. 9, no. 6, pp. 177 –180, June 2002.

[3] M. D. Plumbley, "Algorithms for non-negative independent component analysis," 2002, Submitted for publication.

[4] L. Parra, C. Spence, P. Sajda, A. Ziehe, and K.-R. Müller, "Unmixing hyperspectral data," in *Advances in Neural Information Processing Systems 12 (Proc. NIPS*99)*. 2000, pp. 942–948, MIT Press.

[5] J. S. Lee, D. D. Lee, S. Choi, and D. S. Lee, "Application of non-negative matrix factorization to dynamic positron emission tomography," in *Proceedings of the International Conference on Independent Component Analysis and Signal Separation (ICA2001), San Diego, California*, T.-W. Lee, T.-P. Jung, S. Makeig, and T. J. Sejnowski, Eds., December 9-13 2001, pp. 629–632.

[6] S. Tsuge, M. Shishibori, S. Kuroiwa, and K. Kita, "Dimensionality reduction using non-negative matrix factorization for information retrieval," in *IEEE International Conference on Systems, Man, and Cybernetics*, Tucson, AZ, USA, 7-10 October 2001, pp. 960 – 965 vol.2.

[7] R. C. Henry, "Multivariate receptor models–current practice and future trends," *Chemometrics and Intelligent Laboratory Systems*, vol. 60, no. 1-2, pp. 43–48, Jan. 2002.

[8] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, pp. 111–126, 1994.

[9] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 21 October 1999.

[10] K. I. Diamantaras and S. Y. Kung, *Principal Component Neural Networks: Theory and Applications*, Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, Inc., 1996.

[11] E. Oja, "The nonlinear PCA learning rule in independent component analysis," *Neurocomputing*, vol. 17, no. 1, pp. 25–46, 1997.

[12] L. Xu, "Least mean square error reconstruction principle for self-organizing neural nets," *Neural Networks*, vol. 6, pp. 627–648, 1993.

[13] E. Oja, "Nonlinear PCA criterion and maximum likelihood in independent component analysis," in *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, Aussois, France, 1999, pp. 143–148.

[14] E. Oja, *Subspace Methods of Pattern Recognition*, Research Studies Press, England, and Wiley, USA, 1983.

[15] J. Karhunen and J. Joutsensalo, "Generalizations of principal component analysis, optimization problems, and neural networks," *Neural Networks*, vol. 8, no. 4, pp. 549–562, 1995.

[16] A. Cichocki, W. Kasprzak, and S.-I. Amari, "Neural network approach to blind separation and enhancement of images," in *Signal Processing VIII: Theories and Applications*, G. Ramponi et al., Eds. 1996, vol. I, pp. 579–582, EURASIP/LINT Publ., Trieste, Italy.