LETTER Communicated by Sebastian Seung

Blind Separation of Positive Sources by Globally Convergent Gradient Search

Erkki Oja
erkki.oja@hut.fi
Neural Networks Research Centre, Helsinki University of Technology, 02015 HUT, Finland

Mark Plumbley
mark.plumbley@elec.qmul.ac.uk
Department of Electrical Engineering, Queen Mary, University of London, London E1 4NS, U.K.

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 column 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 letter, we consider the task of independent component analysis when the independent sources are known to be nonnegative and well grounded, which means that they have a nonzero pdf in the region of zero. It can be shown that in this case, the solution method is basically very simple: an orthogonal rotation of the whitened observation vector into nonnegative outputs will give a positive permutation of the original sources. We propose a cost function whose minimum coincides with nonnegativity and derive the gradient algorithm under the whitening constraint, under which the separating matrix is orthogonal. We further prove that in the Stiefel manifold of orthogonal matrices, the cost function is a Lyapunov function for the matrix gradient flow, implying global convergence. Thus, this algorithm is guaranteed to find the nonnegative well-grounded independent sources. The analysis is complemented by a numerical simulation, which illustrates the algorithm.

1 Introduction

The problem of independent component analysis (ICA) has been studied by many authors in recent years (for a review, see Hyvärinen, Karhunen, & Oja, 2001; Amari & Cichocki, 2002). In the simplest form of ICA, we assume that we have a sequence of observations $\{\mathbf{x}(k)\}$, which are samples of a random

 © 2004 Massachusetts Institute of Technology

observation vector $\mathbf{x}$ generated according to

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{1.1}$$

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 that 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 approximations by higher-order cumulants such as kurtosis. The observations $\mathbf{x}$ are usually assumed to be zero mean, or transformed to be so, and are commonly prewhitened 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.

This basic ICA model can be considered to be solved, with a multitude of practical algorithms and software. However, if one makes some further assumptions that restrict or extend the model, then there is still ground for new analysis and solution methods. One such assumption is positivity or nonnegativity of the sources and perhaps the mixing coefficients. Nonnegativity is a natural condition for many real-world applications, for example, in the analysis of images (Parra, Spence, Sajda, Ziehe, & Müller, 2000; Lee, Lee, Choi, & Lee, 2001), text (Tsuge, Shishibori, Kuroiwa, & Kita, 2001) or air quality (Henry, 2002). The constraint of nonnegative sources, perhaps with an additional constraint of nonnegativity on the mixing matrix $\mathbf{A}$, is often known as *positive matrix factorization* (Paatero & Tapper, 1994) or *nonnegative matrix factorization* (Lee & Seung, 1999). A nonnegativity constraint has been suggested for a number of other neural network models too (Xu, 1993; Fyfe, 1994; Harpur, 1997; Charles & Fyfe, 1998). We refer to the combination of nonnegativity and independence assumptions on the sources as *nonnegative independent component analysis*.

Recently, one of us considered the nonnegativity assumption on the sources (Plumbley, 2002, 2003) and introduced an alternative way of approaching the ICA problem, as follows. We call a source s_i *nonnegative* if $\Pr(s_i < 0) = 0$, and such a source will be called *well grounded* if $\Pr(s_i < \delta) > 0$ for any $\delta > 0$, that is, that s_i has nonzero pdf all the way down to zero. The following key result was proven (Plumbley, 2002):

Theorem 1. *Suppose that* $\mathbf{s}$ *is a vector of nonnegative well-grounded independent unit-variance sources* s_i, $i = 1, \ldots, n$, *and* $\mathbf{y} = \mathbf{U}\mathbf{s}$ *where* $\mathbf{U}$ *is a square orthonormal rotation, that is,* $\mathbf{U}^T\mathbf{U} = \mathbf{I}$. *Then* $\mathbf{U}$ *is a permutation matrix, that is,*

the elements y_j of $\mathbf{y}$ are a permutation of the sources s_i, if and only if all y_j are nonnegative.

Actually, the requirement for independence of the sources in theorem 1 is a technicality that simplifies the proof but could be relaxed to second-order independence or uncorrelatedness. The proof would be lengthy, however, and we wished to rely on the existing theorem 1 in this letter.

The result of theorem 1 can be used for a simple solution of the nonnegative 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 prewhitened 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 nonnegative. This brings 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 nonnegative. The sign ambiguity present in usual ICA vanishes here.

Plumbley (2002) further suggested that a suitable cost function for finding the rotation could be constructed as follows. Suppose we have an output truncated at zero, $\mathbf{y}^+ = (y_1^+, \ldots, y_n^+)$ with $y_i^+ = \max(0, y_i)$, and we construct a reestimate of $\mathbf{z} = \mathbf{W}^T\mathbf{y}$ given by $\hat{\mathbf{z}} = \mathbf{W}^T\mathbf{y}^+$. Then a suitable cost function would be given by

$$J(\mathbf{W}) = E\{\|\mathbf{z} - \hat{\mathbf{z}}\|^2\} = E\{\|\mathbf{z} - \mathbf{W}^T\mathbf{y}^+\|^2\}, \tag{1.2}$$

because obviously its value will be zero if $\mathbf{W}$ is such that all the y_i are positive, or $\mathbf{y} = \mathbf{y}^+$. We considered the minimization of this cost function by various numerical algorithms in Plumbley (2003) and Plumbley and Oja (2004). In Plumbley (2003), explicit axis rotations as well as geodesic search over the Stiefel manifold of orthogonal matrices were used. In Plumbley and Oja (2004), the cost function 1.2 was taken as a special case of nonlinear PCA, for which an algorithm was earlier suggested by Oja (1997). However, a rigorous convergence proof for the nonlinear PCA method could not be constructed except in some special cases. The general convergence seems to be a very challenging problem.

The new key result shown in this article is that the cost function 1.2 has very desirable properties. In the Stiefel manifold of rotation matrices, the function has no local minima, and it is a Lyapunov function for its gradient matrix flow. A gradient algorithm, suggested in the following, is therefore monotonically converging and is guaranteed to find the absolute minimum of the cost function. The minimum is zero, giving positive components y_i, which by theorem 1 must be a positive permutation of the original unknown sources s_j. Some preliminary results along these lines were given in Oja and Plumbley (2003).

In the next section, 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. In section 3, we consider the cost function

1.2 in more detail. Section 4 gives a gradient algorithm, whose monotonical global convergence is proven. Section 5 relates this orthogonalized algorithm to the nonorthogonal "nonlinear PCA" learning rule previously introduced by one of the authors by illustrating both algorithms in a pictorial example. Finally, section 6 gives some conclusions.

2 Prewhitening 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

$$\mathbf{z} = \mathbf{V}\mathbf{x}, \tag{2.1}$$

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

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

$\Sigma_{\mathbf{x}}$ is normally estimated from the sample covariance (Hyvärinen et al., 2001). Note that for nonnegative ICA, we do not remove the mean of the data, since this would lose information about the nonnegativity of the sources (Plumbley, 2002).

Suppose that our sources s_j have unit variance, such that $\Sigma_{\mathbf{s}} = \mathbf{I}_n$, and let $\mathbf{U} = \mathbf{V}\mathbf{A}$ be the $\mathbf{s}$-to-$\mathbf{z}$ transform. Then $\mathbf{I}_n = \Sigma_{\mathbf{z}} = \mathbf{U}\Sigma_{\mathbf{s}}\mathbf{U}^T = \mathbf{U}\mathbf{U}^T$ 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 nonnegative data in two dimensions. Whitening has succeeded in making the axes of the original sources orthogonal to each other (see Figure 1b), but there is a remaining orthonormal rotation ambiguity. A typical ICA algorithm might search for a rotation that makes the resulting outputs as nongaussian as possible, for example, by finding an extremum of kurtosis, since any sum of independent random variables will make the result "more gaussian" (Hyvärinen et al., 2001).

However, Figure 1 immediately suggests another approach: we should search for a rotation where all the data fit 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. Note also that after this rotation, the two sources s_1 and s_2 are indeed independent, even

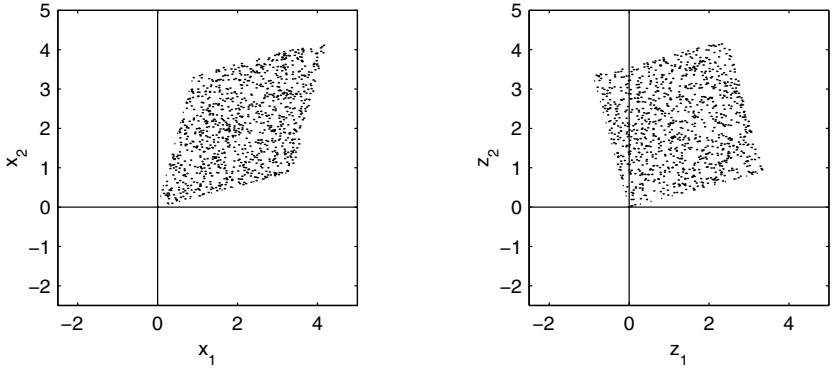

Figure 1: Original data (a) are whitened (b) to remove second-order correlations.

if they are not zero mean, because for their densities, it holds that the joint density (in this case, uniform in a square) is the product of the marginal (uniform) densities.

3 The Cost Function for Nonnegative BSS

In the following, we show that the minimum of the cost function 1.2 in the set of orthogonal (rotation) matrices will give the sources. For brevity of notation, let us denote the truncation nonlinearity by

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

which is zero for negative y_i and y_i otherwise.

We are now ready to state and prove theorem 2:

Theorem 2. *Assume the n-element random vector* $\mathbf{z}$ *is a whitened linear mixture of nonnegative well-grounded independent unit variance sources* $s_1, \ldots, s_n$, *and* $\mathbf{y} = \mathbf{W}\mathbf{z}$ *with* $\mathbf{W}$ *constrained to be a square orthogonal matrix. If* $\mathbf{W}$ *is obtained as the minimum of the cost function 1.2, rewritten as*

$$J(\mathbf{W}) = E\|\mathbf{z} - \mathbf{W}^T g^+(\mathbf{W}\mathbf{z})\|^2,$$

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:

$$J(\mathbf{W}) = E\{\|\mathbf{z} - \mathbf{W}^T g^+(\mathbf{W}\mathbf{z})\|^2\} \tag{3.1}$$

$$= E\{\|\mathbf{Wz} - \mathbf{WW}^T g^+(\mathbf{Wz})\|^2\} \tag{3.2}$$
$$= E\{\|\mathbf{y} - g^+(\mathbf{y})\|^2\} \tag{3.3}$$
$$= \sum_{i=1}^{n} E\{[y_i - g^+(y_i)]^2\} \tag{3.4}$$
$$= \sum_{i=1}^{n} E\{\min(0, y_i)^2\} \tag{3.5}$$
$$= \sum_{i=1}^{n} E\{y_i^2 | y_i < 0\} P(y_i < 0). \tag{3.6}$$

This is always nonnegative and becomes zero if and only if each y_i is nonnegative with probability one. Thus, if $\mathbf{W}$ is obtained as the minimum of the cost function, then the elements of $\mathbf{y}$ will be nonnegative.

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

4 A Converging Gradient Algorithm

Theorem 2 leads us naturally to consider the use of a gradient algorithm for minimizing equation 3.1 under the orthogonality constraint. In order to derive the gradient, let us first write equation 3.1 in the simple form, equation 3.5:

$$J(\mathbf{w}_1, \ldots, \mathbf{w}_n) = \sum_{i=1}^{n} E\{\min(0, y_i)^2\}, \; y_i = \mathbf{w}_i^T \mathbf{z}, \tag{4.1}$$

where the vectors $\mathbf{w}_i^T$ are the rows of matrix $\mathbf{W}$ and thus satisfy

$$\mathbf{w}_i^T \mathbf{w}_j = \delta_{ij}. \tag{4.2}$$

The gradient with respect to one of vectors $\mathbf{w}_i$ is straightforward:

$$\frac{\partial J}{\partial \mathbf{w}_i} = E\left\{ \frac{\min(0, y_i)^2}{\partial y_i} \frac{\partial y_i}{\partial \mathbf{w}_i} \right\} = 2E\{\min(0, y_i)\mathbf{z}\}. \tag{4.3}$$

A possible way to do the constrained minimization by gradient descent is to divide each descent step into two parts: a step in the direction of the unconstrained gradient, followed by a consequent projection of the new point onto the constraint set 4.2. For vectors $\mathbf{w}_i$, this gives the update rule

$$\tilde{\mathbf{w}}_i = \mathbf{w}_i - \gamma \frac{\partial J}{\partial \mathbf{w}_i} \tag{4.4}$$
$$= \mathbf{w}_i - 2\gamma E\{\min(0, y_i)\mathbf{z}\}, \tag{4.5}$$

which is the unconstrained gradient descent step with step size γ, and

$$\mathbf{W} = (\tilde{\mathbf{W}}\tilde{\mathbf{W}}^T)^{-1/2}\tilde{\mathbf{W}}, \tag{4.6}$$

which is the projection onto the constraint set of orthonormal $\mathbf{w}_i^T$ vectors, because equation 4.6 implies $\mathbf{W}\mathbf{W}^T = \mathbf{I}$. Obviously, matrix $\tilde{\mathbf{W}}$ is the one with vectors $\tilde{\mathbf{w}}_i{}^T$ as its rows.

This is just one of the possibilities for performing orthonormalization; another alternative would be the Gram-Schmidt algorithm. However, we prefer a symmetrical orthonormalization as there is no reason to order the $\mathbf{w}_i$ vectors in any way.

In matrix form, equation 4.5 reads

$$\tilde{\mathbf{W}} = \mathbf{W} - 2\gamma E\{\mathbf{f}\mathbf{z}^T\}, \tag{4.7}$$

where $\mathbf{f} = \mathbf{f}(\mathbf{y})$ is the column vector with elements $\min(0, y_i)$. This is the gradient descent step. Now, to derive the projection in equation 4.6, we have $\tilde{\mathbf{W}}\tilde{\mathbf{W}}^T = \mathbf{W}\mathbf{W}^T + 4\gamma^2 E\{\mathbf{f}\mathbf{z}^T\}E\{\mathbf{z}\mathbf{f}^T\} - 2\gamma E\{\mathbf{W}\mathbf{z}\mathbf{f}^T + \mathbf{f}\mathbf{z}^T\mathbf{W}^T\} = \mathbf{I} - 2\gamma E\{\mathbf{W}\mathbf{z}\mathbf{f}^T + \mathbf{f}\mathbf{z}^T\mathbf{W}^T\} + O(\gamma^2)$. Thus, assuming γ small, $(\tilde{\mathbf{W}}\tilde{\mathbf{W}}^T)^{-1/2} = \mathbf{I} + \gamma E\{\mathbf{W}\mathbf{z}\mathbf{f}^T + \mathbf{f}\mathbf{z}^T\mathbf{W}^T\} + O(\gamma^2)$, and finally $(\tilde{\mathbf{W}}\tilde{\mathbf{W}}^T)^{-1/2}\tilde{\mathbf{W}} = \mathbf{W} - 2\gamma E\{\mathbf{f}\mathbf{z}^T\} + \gamma E\{\mathbf{W}\mathbf{z}\mathbf{f}^T\mathbf{W} + \mathbf{f}\mathbf{z}^T\} + O(\gamma^2) = \mathbf{W} - \gamma E\{\mathbf{f}\mathbf{z}^T\} + \gamma E\{\mathbf{y}\mathbf{f}^T\mathbf{W}\} + O(\gamma^2)$. Omitting $O(\gamma^2)$, the change from $\mathbf{W}$ at the previous step to the new $\mathbf{W} = (\tilde{\mathbf{W}}\tilde{\mathbf{W}}^T)^{-1/2}\tilde{\mathbf{W}}$ at the next step is therefore $\Delta\mathbf{W} = -\gamma E\{\mathbf{f}\mathbf{z}^T - \mathbf{y}\mathbf{f}^T\mathbf{W}\}$, which can be further written in the form

$$\Delta\mathbf{W} = -\gamma E\{\mathbf{f}\mathbf{y}^T - \mathbf{y}\mathbf{f}^T\}\mathbf{W}. \tag{4.8}$$

Substituting $\min(0, y_i)$ for the elements of $\mathbf{f}(\mathbf{y})$ gives the gradient descent algorithm for the constrained problem.

The form of the update rule 4.8 is the same as that derived by Cardoso and Laheld (1996) (although with different notation) for minimizing a general contrast under the orthogonality constraint. However, they are projecting the unconstrained gradient onto the space of skew-symmetric matrices. It seems that from our derivation, higher-order terms with respect to the step size γ and thus a more accurate projection could be more easily obtained by continuing the series expansion.

The skew-symmetric form of the matrix $\mathbf{f}\mathbf{y}^T - \mathbf{y}\mathbf{f}^T$ in equation 4.8 ensures that $\mathbf{W}$ tends to stay orthogonal from step to step, although to fully guarantee orthogonality in a discrete-time gradient algorithm, an explicit orthonormalization of the rows of $\mathbf{W}$ should be done from time to time.

Instead of analyzing this learning rule directly, let us look at the averaged differential equation corresponding to the discrete-time algorithm 4.8. It becomes

$$\frac{d\mathbf{W}}{dt} = -\mathbf{M}\mathbf{W}, \tag{4.9}$$

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

$$\mu_{ij} = E\{\min(0, y_i)y_j - y_i \min(0, y_j)\}. \tag{4.10}$$

Note that $\mathbf{M}$ is a nonlinear function of the solution $\mathbf{W}$, because $\mathbf{y} = \mathbf{W}\mathbf{z}$. Yet we can formally write the solution of equation 4.9 as

$$\mathbf{W}(t) = \exp\left[-\int_0^t \mathbf{M}(s)ds\right]\mathbf{W}(0). \tag{4.11}$$

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

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

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 equation 4.9 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 $\mathbf{M}\mathbf{W} = 0$, which is equivalent to $\mathbf{M} = 0$ because of the orthogonality of $\mathbf{W}$. 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 equation 4.10 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 equation 4.10 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 $J(\mathbf{W})$. 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 ordinary differential equation (ODE) will converge to it.

The case when all the y_i are negative corresponds to the maximum value of $J(\mathbf{W})$, 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 equation 4.11.

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 equation 3.1 is in fact a Lyapunov function for the matrix flow 4.9; it is strictly decreasing always when $\mathbf{W}$ changes according to the ODE 4.9, except at the stationary points. Let us prove this in the following.

Theorem 3. *If* $\mathbf{W}$ *follows the ODE 4.9, then* $\frac{dJ(\mathbf{W})}{dt} < 0$, *except at the point when all* y_i *are nonnegative or all are nonpositive.*

Proof. Consider the ith term in the sum $J(\mathbf{W})$, given in equation 3.5. Denoting it by e_i, we have $e_i = E\{\min(0, y_i)^2\}$ whose derivative with respect to y_i is $2E\{\min(0, y_i)\}$. If $\mathbf{w}_i^T$ is the ith row of matrix $\mathbf{W}$, then $y_i = \mathbf{w}_i^T\mathbf{z}$. Thus,

$$\frac{de_i}{dt} = \frac{de_i}{dy_i}\frac{dy_i}{dt} = 2E\left\{\min(0, y_i)\left(\frac{d\mathbf{w}_i^T}{dt}\mathbf{z}\right)\right\}. \tag{4.12}$$

From the ODE 4.9, we get

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

with μ_{ik} given in equation 4.10. Substituting this in equation 4.12 gives

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

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

$$\frac{dJ(\mathbf{W})}{dt} = \sum_{i=1}^{n}\frac{de_i}{dt} = 2\left[-\sum_{i=1}^{n}\sum_{k=1}^{n}\alpha_{ik}^2 + \sum_{i=1}^{n}\sum_{k=1}^{n}\alpha_{ik}\alpha_{ki}\right].$$

By the Cauchy-Schwartz inequality, this is strictly negative unless $\alpha_{ik} = \alpha_{ki}$ for all i, k, and thus $J(\mathbf{W})$ is decreasing.

We still have to look at the condition that $\alpha_{ik} = \alpha_{ki}$ for all i, k and show that this implies nonnegativity or nonpositivity 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 nonnegative. 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 nonnegative, 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 nonnegative, then all y_k must be nonnegative in the case of symmetrical α_{ij}.

The behavior of the learning rule 4.8 is now well understood. The function $J(\mathbf{W})$ in equation 1.2 is a Lyapunov function for the averaged differential equation 4.9, for all orthogonal matrices $\mathbf{W}$ except for the point with all outputs y_i nonpositive. Therefore, recalling that if $\mathbf{W}(0)$ is an orthogonal matrix, then it is constrained to remain so, $\mathbf{W}(t)$ converges to the minimum of $J(\mathbf{W})$ from almost everywhere in the Stiefel manifold of orthogonal matrices. For a discussion of optimization and learning on the Stiefel manifold, see Edelman, Arias, and Smith (1998) and Fiori (2001). This minimum corresponds to all nonnegative y_i. By theorem 1, these must be a permutation of the original sources s_j which therefore have been found.

The result was proven only for the continuous-time averaged version of the learning rule; the exact connection between this and the discrete-time online algorithm has been clarified in the theory of stochastic approximation (see Oja, 1983). In practice, even if the starting point happened to be the "bad" stationary point in which all y_i are nonpositive, then numerical errors will deviate the solution from this point and the cost function $J(\mathbf{W})$ starts to decrease.

5 Experiments

We illustrate the operation of the algorithm using a blind image separation problem. We use the same images and perfomance measures as in Plumbley and Oja (2004), in which the nonlinear PCA algorithm (Oja, 1997) was used instead to solve the nonnegative ICA problem. As performance measures, we measure a mean squared error e_{MSE}, an orthonormalization error e_{Orth}, and a permutation error e_{Perm}, defined as follows:

$$e_{MSE} = \frac{1}{np} \sum_{k=1}^{p} \|\mathbf{z}_k - \mathbf{W}^T g^+(\mathbf{y}_k)\|^2 \tag{5.1}$$

$$e_{Orth} = \frac{1}{n^2} \|\mathbf{I} - (\mathbf{WVA})^T \mathbf{WVA}\|_F^2 \tag{5.2}$$

$$e_{Perm} = \frac{1}{n^2} \|\mathbf{I} - \mathrm{abs}(\mathbf{WVA})^T \mathrm{abs}(\mathbf{WVA})\|_F^2, \tag{5.3}$$

where abs($\mathbf{M}$) returns the absolute value of each element of $\mathbf{M}$, so that $e_{Perm} = 0$ only for a positive permutation matrix. The parameters have been scaled (by $1/(np)$ or $1/n^2$) to allow more direct comparison between the result of simulations using different values for n and p.

Four image patches of size 252×252 were selected from a set of images of natural scenes and downsampled by a factor of 4 in both directions to yield 63×63 pixel images. Each of the $n = 4$ images was treated as one source, with its pixel values representing the $p = 63 \times 63 = 3969$ samples. The source image values were shifted to have a minimum of zero to ensure they were well grounded, and the images were scaled to ensure they were all unit variance. After scaling, the source covariance matrix was found to be

$$\overline{\mathbf{s}\mathbf{s}^T} - \bar{\mathbf{s}}\bar{\mathbf{s}}^T = \begin{pmatrix} 1.000 & 0.074 & -0.003 & 0.050 \\ 0.074 & 1.000 & -0.071 & 0.160 \\ -0.003 & -0.071 & 1.000 & 0.130 \\ 0.050 & 0.160 & 0.130 & 1.000 \end{pmatrix}, \tag{5.4}$$

giving an acceptably small covariance between the images. A mixing matrix $\mathbf{A}$ was generated randomly and used to construct $\mathbf{x} = \mathbf{A}\mathbf{s}$.

For the algorithm, the demixing matrix $\mathbf{W}$ was initialized to the identity matrix, ensuring initial orthogonality of $\mathbf{W}$. Instead of algorithm 4.8, the theoretical expectation was replaced by a batch update method: denoting by $\mathbf{X}$ the observation matrix having all the data vectors $\mathbf{x}(1), \ldots, \mathbf{x}(T)$ as its columns and defining the matrix of outputs as $\mathbf{Y} = \mathbf{W}\mathbf{X}$, we update $\mathbf{W}$ with the incremental change,

$$\Delta\mathbf{W} = -\mu \frac{1}{p}\{\mathbf{f}(\mathbf{Y})\mathbf{Y}^T - \mathbf{Y}\mathbf{f}(\mathbf{Y})^T\}\mathbf{W}. \tag{5.5}$$

A constant update factor of $\mu = 0.03$ was used, and $\mathbf{W}$ was renormalized to an orthogonal matrix after each step using equation 4.6.

Figure 2 shows the performance of learning over 2×10^4 steps, with Figure 3 showing the original, mixed, and separated images and their histograms. After 2×10^4 iteration steps, in which each step is one update following presentation of the batch of 3969 samples (1610s/27min of CPU time on an 850 MHz Pentium III), the source-to-output matrix $\mathbf{WVA}$ was found to be

$$\mathbf{WVA} = \begin{pmatrix} \mathbf{1.002} & 0.015 & -0.040 & 0.026 \\ -0.099 & 0.067 & -0.111 & \mathbf{1.007} \\ -0.016 & \mathbf{1.009} & -0.089 & 0.018 \\ 0.004 & -0.056 & \mathbf{1.021} & -0.058 \end{pmatrix}, \tag{5.6}$$

with $e_{MSE} = 6.07 \times 10^{-5}$, $e_{Orth} = 8.88 \times 10^{-3}$, and $e_{Perm} = 1.07 \times 10^{-2}$.

The mean squared error and orthogonalization error are slightly better than for the nonnegative PCA algorithm, which were 9.30×10^{-5} and 9.02×10^{-3}, respectively, for the same number of iterations. The final permutation error for the same number of iterations is also smaller than the value obtained with the nonnegative PCA algorithm, which was 1.68×10^{-2}.

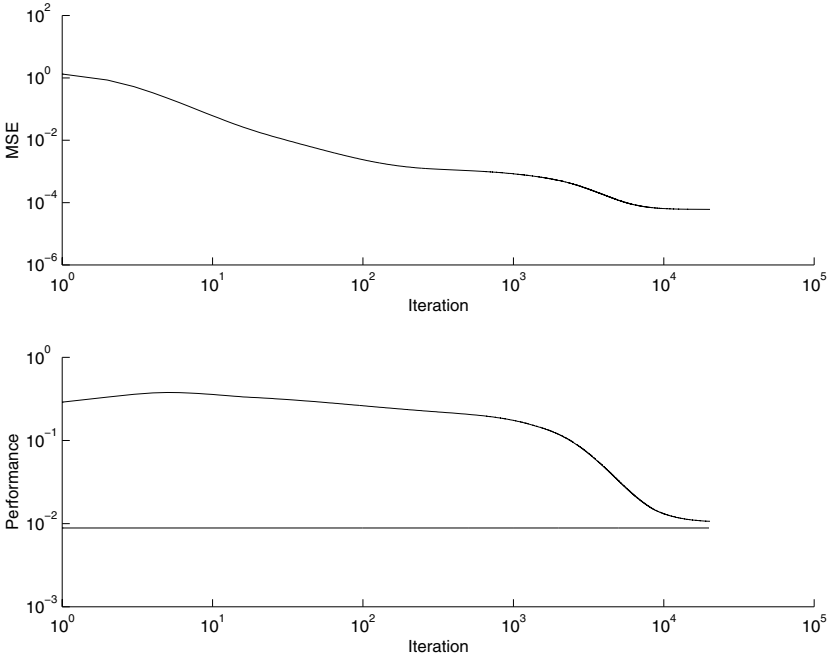

Figure 2: Simulation results on image data. Performance in the lower graph is measured as distance from permutation e_{Perm} (upper curve) and distance from orthogonality e_{Orth} (lower curve). See the text for definitions.

The lower bound on e_{Orth} and e_{Perm} is determined by the accuracy of the prewhitening stage: recall that prewhitening is estimated from the statistics of the input data, without having any access to the original mixing matrix $\mathbf{A}$. Calculating the equivalent error in $\mathbf{VA}$ from orthonormality,

$$e_{White} = \frac{1}{n^2}\|\mathbf{I} - (\mathbf{VA})^T(\mathbf{VA})\|_F^2, \tag{5.7}$$

we find $e_{White} = 8.88 \times 10^{-3} = e_{Orth}$ to within the machine accuracy (i.e., $|e_{Orth} - e_{White}| < 10^{-15}$) as we might expect, since $\mathbf{W}$ is orthonormalized at each iteraction, so $\mathbf{WW}^T = \mathbf{I}$. Therefore, the e_{Perm} is 20.3% above its lower bound for this algorithm, compared to 80.3% for the nonnegative PCA algorithm.

6 Discussion

We have considered the problem of nonnegative ICA, that is, independent component analysis where the sources are known to be nonnegative. Else-

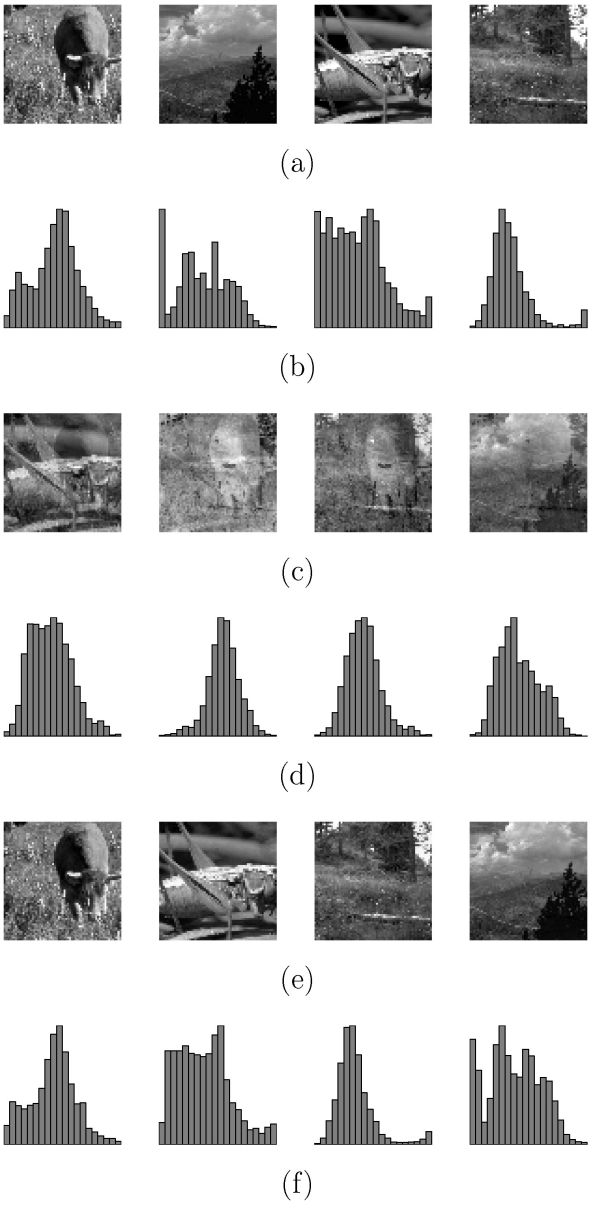

Figure 3: Images and histograms for the image separation task, showing (a) the original source images (b) and their histograms, (c, d) the mixed images and their histograms, and (e, f) the separated images and their histograms.

where, one of us introduced algorithms to solve this based on the use of orthogonal rotations, related to Stiefel manifold approaches (Plumbley, 2003).

In this article, we considered a gradient-based algorithm operating on prewhitened data, related to the "nonlinear PCA" algorithms investigated by one of the authors (Oja, 1997, 1999). We refer to these algorithms, which use a truncation nonlinearity, as nonnegative PCA algorithms. By theoretical analysis of algorithm 4.8, we showed the key result of the article: asymptotically, as the learning rate is very small, the algorithm is guaranteed to find a permutation of the well-grounded nonnegative sources. Such a global convergence result is rather unique in ICA gradient methods. The convergence was experimentally verified for a small learning rate using a set of positive images as sources, with a random mixing matrix.

Acknowledgments

Patrik Hoyer kindly supplied the images used in section 5. Part of this work was undertaken while M.P. was visiting the Neural Networks Research Centre at the Helsinki University of Technology, supported by a Leverhulme Trust Study Abroad Fellowship. 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.

References

Amari, S.-I., & Cichocki, A. (2002). *Adaptive blind signal and image processing*. New York: Wiley.

Cardoso, J.-F., & Laheld, B. H. (1996). Equivariant adaptive source separation. *IEEE Transactions on Signal Processing, 44*(12), 3017–3030.

Charles, D., & Fyfe, C. (1998). Modelling multiple-cause structure using rectification constraints. *Network: Computation in Neural Systems, 9*, 167–182.

Edelman, A., Arias, T. A., & Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl., 20*(2), 303–353.

Fiori, S. (2001). A theory for learning by weight flow on Stiefel-Grassman manifold. *Neural Computation, 13*, 1625–1647.

Fyfe, C. (1994). Positive weights in interneurons. In G. Orchard (Ed.), *Neural computing: Research and applications II. Proceedings of the Third Irish Neural Networks Conference, Belfast, Northern Ireland, 1–2 Sept 1993* (pp. 47–58). Belfast, NI: Irish Neural Networks Association.

Harpur, G. F. (1997). *Low entropy coding with unsupervised neural networks*. Unpublished doctoral dissertation, Cambridge University.

Henry, R. C. (2002). Multivariate receptor models—current practice and future trends. *Chemometrics and Intelligent Laboratory Systems, 60*(1–2), 43–48.

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. New York: Wiley.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by nonnegative matrix factorization. *Nature, 401*, 788–791.

Lee, J. S., Lee, D. D., Choi, S., & Lee, D. S. (2001). Application of nonnegative matrix factorization to dynamic positron emission tomography. In T.-W. Lee, T.-P. Jung, S. Makeig, & T. J. Sejnowski (Eds.), *Proceedings of the International Conference on Independent Component Analysis and Signal Separation (ICA2001), San Diego, California* (pp. 629–632). San Diego, CA: Institute of Neural Computation, University of California, San Diego.

Oja, E. (1983). *Subspace methods of pattern recognition*. Baldock, U.K.: Research Studies Press, and New York: Wiley.

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

Oja, E. (1999). Nonlinear PCA criterion and maximum likelihood in independent component analysis. In *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)* (pp. 143–148). Aussois, France: ICA 1999 Organizing Committee, Institut National Polytechnique de Grenoble, France.

Oja, E., & Plumbley, M. D. (2003). Blind separation of positive sources using nonnegative ICA. In *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'03)*. Kyoto, Japan: ICA 2003 Organizing Committee, NTT Communication Science Laboratories.

Paatero, P., & Tapper, U. (1994). Positive matrix factorization: A nonnegative factor model with optimal utilization of error estimates of data values. *Environmetrics, 5*, 111–126.

Parra, L., Spence, C., Sajda, P., Ziehe, A., & Müller, K.-R. (2000). Unmixing hyperspectral data. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in neural information processing systems, 12* (pp. 942–948). Cambridge, MA: MIT Press.

Plumbley, M. D. (2002). Conditions for nonnegative independent component analysis. *IEEE Signal Processing Letters, 9*(6), 177–180.

Plumbley, M. D. (2003). Algorithms for non-negative independent component analysis. *IEEE Transactions on Neural Networks, 14*(3), 534–543.

Plumbley, M. D., & Oja, E. (2004). A "non-negative PCA" algorithm for independent component analysis. *IEEE Transactions on Neural Networks, 15*(1), 66–76.

Tsuge, S., Shishibori, M., Kuroiwa, S., & Kita, K. (2001). Dimensionality reduction using non-negative matrix factorization for information retrieval. In *IEEE International Conference on Systems, Man, and Cybernetics* (Vol. 2, pp. 960–965). Piscataway, NJ: IEEE.

Xu, L. (1993). Least mean square error reconstruction principle for self-organizing neural-nets. *Neural Networks, 6*(5), 627–648.

Received July 9, 2003; accepted February 5, 2004.