A "nonnegative PCA" algorithm for independent component analysis

Mark D. Plumbley and Erkki Oja

IEEE Transactions on Neural Networks, 15(1), 66-76, Jan 2004

Abstract

We consider the task of independent component analysis when the independent sources are known to be nonnegative and well-grounded, so that they have a nonzero probability density function (pdf) in the region of zero. We propose the use of a "nonnegative principal component analysis (nonnegative PCA)" algorithm, which is a special case of the nonlinear PCA algorithm, but with a rectification nonlinearity, and we conjecture that this algorithm will find such nonnegative well-grounded independent sources, under reasonable initial conditions. While the algorithm has proved difficult to analyze in the general case, we give some analytical results that are consistent with this conjecture and some numerical simulations that illustrate its operation.

Index Terms

independent component analysis learning (artificial intelligence) matrix decomposition principal component analysis independent component analysis nonlinear principal component analysis nonnegative PCA algorithm nonnegative matrix factorization nonzero probability density function rectification nonlinearity subspace learning rule

A "Nonnegative PCA" Algorithm for Independent Component Analysis

Mark D. Plumbley, *Member, IEEE,* and Erkki Oja, *Fellow, IEEE*

***Abstract*—We consider the task of independent component analysis when the independent sources are known to be nonnegative and well-grounded, so that they have a nonzero probability density function (pdf) in the region of zero. We propose the use of a "nonnegative principal component analysis (nonnegative PCA)" algorithm, which is a special case of the nonlinear PCA algorithm, but with a rectification nonlinearity, and we conjecture that this algorithm will find such nonnegative well-grounded independent sources, under reasonable initial conditions. While the algorithm has proved difficult to analyze in the general case, we give some analytical results that are consistent with this conjecture and some numerical simulations that illustrate its operation.**

***Index Terms*—Independent component analysis (ICA), nonlinear principal component analysis (nonlinear PCA), nonnegative matrix factorization, subspace learning rule.**

I. Introduction

THE problem of independent component analysis (ICA) has been studied by many authors in recent years (see, e.g., [1]). In the simplest form of ICA we assume that we have a sequence of observations $\{\mathbf{x}_k\}$ with each 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*) 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 $\mathbf{s}$ with just a sign ambiguity.

Common methods for ICA involve higher order cumulants such as kurtosis or the use of autocorrelation differences between the sources. The observations $\mathbf{x}$ are often assumed to be zero-mean, or transformed to be so, and are commonly prewhitened by some matrix $\mathbf{V}$ giving $\mathbf{z} = \mathbf{V}\mathbf{x}$ so that $E\{\mathbf{z}\mathbf{z}^T\} = \mathbf{I}$ before an optimization algorithm is applied to $\mathbf{z}$.

Recently, one of the current authors considered an additional assumption on the sources: that they are *nonnegative* as well as independent [2], [3]. Nonnegativity is a natural condition for many real-world applications, for example in the analysis of images [4], [5], text [6], or air quality [7]. Neural networks have been suggested that impose a nonnegativity constraint on the outputs [8]–[10] or weights [11], [12]. 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* [13] or *nonnegative matrix factorization* [14]. We refer to the combination of nonnegativity and independence assumptions on the sources as *nonnegative independent component analysis*.

Nonnegativity of sources can provide us with 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 $\delta > 0$, i.e., that s_i has nonzero pdf all the way down to zero. Suppose that $\mathbf{s}$ is a vector of real nonnegative well-grounded independent unit-variance sources and $\mathbf{y} = \mathbf{U}\mathbf{s}$, where $\mathbf{U}$ is an orthonormal rotation, i.e., $\mathbf{U}^T\mathbf{U} = \mathbf{I}$. Then $\mathbf{U}$ is a permutation matrix, i.e., $\mathbf{y}$ is a permutation of the source vector $\mathbf{s}$, if and only if $\mathbf{y}$ is nonnegative [2]. It therefore suffices to find an orthonormal matrix $\mathbf{W}$ for which $\mathbf{y} = \mathbf{W}\mathbf{z}$ is nonnegative. This brings the additional benefit over other ICA methods that we know that, if successful, we have positive permutation of the sources, since both the $\mathbf{s}$ and $\mathbf{y}$ are nonnegative.

This has a connection with principal component analysis (PCA). Recall that, given an n dimensional vector $\mathbf{x}$, its k-dimensional principal component subspace can be found by minimizing the representation error

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

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 $\mathbf{x}$ [15].

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 $g(\mathbf{W}\mathbf{x})$, 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 [16]. Let us write

Manuscript received October 31, 2002; revised June 4, 2003. M. D. Plumbley was supported by a three-month Study Abroad Fellowship from Leverhulme Trust, which allowed him undertake this work at Helsinki University of Technology, Helsinki, Finland. This work was supported by the Centre of Excellence Program of the Academy of Finland, Project New Information Processing Principles, 44886, and in part by Grant GR/R54620 from the U.K. Engineering and Physical Sciences Research Council.

M. D. Plumbley is with the Department of Electronic Engineering, Queen Mary, University of London, London E1 4NS, U.K. (e-mail: mark.plumbley@elec.qmul.ac.uk).

E. Oja is with the Laboratory of Computer and Information Sciences, Helsinki University of Technology, 02015 HUT, Finland (e-mail: erkki.oja@hut.fi).

Digital Object Identifier 10.1109/TNN.2003.820672

this nonlinear MSE criterion, first introduced by Xu [8], for the whitened vector $\mathbf{z}$ instead

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

For the whitened case, we can restrict the matrix $\mathbf{W}$ to be square orthogonal ($\mathbf{W}\mathbf{W}^T = \mathbf{W}^T\mathbf{W} = \mathbf{I}$) and it can be shown that [16]

$$e_{\mathrm{MSE}} = E\|\mathbf{y} - g(\mathbf{y})\|^2 = \sum_{i=1}^{n} E\{[y_i - g(y_i)]^2\} \tag{3}$$

with $\mathbf{y} = \mathbf{W}\mathbf{z}$. This was related to some ICA cost functions in [16].

Now, this new criterion has a very simple connection to the problem of finding a nonnegative vector $\mathbf{y}$. Let us use the rectification nonlinearity

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

which is zero for negative y_i and equal to y_i otherwise. Then, (3) becomes

$$\sum_{i=1}^{n} E\{[y_i - g_+(y_i)]^2\} = \sum_{i=1}^{n} E\left\{y_i^2 \,|\, y_i < 0\right\}. \tag{4}$$

This is clearly zero if and only if each y_i is nonnegative with probability one.

To conclude: we are given $\mathbf{z}$, which is a whitened linear mixture of nonnegative sources $s_i, \ldots, s_n$ and we set $\mathbf{y} = \mathbf{W}\mathbf{z}$ with $\mathbf{W}$ constrained to be a square orthogonal matrix. If $\mathbf{W}$ is obtained as the minimum of $E\|\mathbf{z} - \mathbf{W}^T g_+(\mathbf{W}\mathbf{z})\|^2$ where $g_+(\mathbf{y}) = (g_+(y_1), \ldots, g_+(y_n))$, then the elements of $\mathbf{y}$ will be a permutation of the original sources s_i.

This leads us naturally to consider the use of the parameter update algorithm $\mathbf{W} \leftarrow \mathbf{W} + \Delta\mathbf{W}$ with $\Delta\mathbf{W}$ given by

$$\Delta\mathbf{W} = \mu g_+(\mathbf{y})(\mathbf{z} - \mathbf{W}^T g_+(\mathbf{y}))^T \tag{5}$$

where μ is a small update factor, and $\mathbf{z} = \mathbf{V}\mathbf{x}$ is pre-whitened (but nonzero-mean) input data. This is a special case of the *nonlinear PCA (subspace) rule* [17], [18]

$$\Delta\mathbf{W} = \mu g(\mathbf{y})(\mathbf{z} - \mathbf{W}^T g(\mathbf{y}))^T \tag{6}$$

with $g(\mathbf{y}) = g_+(\mathbf{y})$. We shall refer to (5) as the *nonnegative PCA* rule.

To strictly clarify the connection of this learning rule to the nonlinear MSE criterion (2), with $g = g_+$ and $\mathbf{W}$ square orthogonal, the matrix gradient of the criterion (4) is [19]

$$\frac{\partial e_{\mathrm{MSE}}}{\partial \mathbf{W}} = -E\{\mathbf{F}(\mathbf{y})\mathbf{W}\mathbf{r}\mathbf{z}^T + g_+(\mathbf{y})\mathbf{r}^T\} \tag{7}$$

where $\mathbf{r} = \mathbf{z} - \mathbf{W}^T g_+(\mathbf{W}\mathbf{z})$ is the representation error and $\mathbf{F}(\mathbf{y}) = \mathrm{diag}(g'_+(y_1), \ldots, g'_+(y_n))$. Looking at the first term on the right-hand side of (7), 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, since if $y_i < 0$, then $g'_+(y_i) = 0$ and if $y_i \geq 0$, then this term becomes $g'_+(y_i)[y_i - y_i] = 0$.

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 (5) is the exact on-line gradient descent rule for the e_{MSE} criterion with orthonormal $\mathbf{W}$. Note that in gradient descent, the negative gradient has to be used.

However, in the analysis above we used the assumption that matrix $\mathbf{W}$ stays orthogonal. This is not strictly true in the gradient algorithm, unless an explicit orthogonalization is performed at each step. The approximation is better, the closer $\mathbf{W}$ is to orthogonality. It will be shown in the following, that stationary points of the approximating gradient algorithm will be orthogonal matrices; thus, asymptotically, the orthogonality assumption holds and the approximating gradient coincides with the exact gradient.

We conjecture that under certain reasonable conditions, algorithm (5) will extract the set of nonnegative independent sources $\mathbf{s}$, modulo a positive permutation ambiguity. It is not hard to see that if $\mathbf{y}$ is always positive, so $g_+(\mathbf{y}) = \mathbf{y}$, then $\Delta\mathbf{W} = 0$ in (5) if $e_{\mathrm{MSE}} = 0$, so the nonnegative PCA rule (5) is stationary when the nonnegative sources have been identified. However, as we shall see in what follows, a full analysis of this algorithm appears to be particularly difficult, even for the case $n = 2$. The analysis given in [18] applies only to odd, twice-differential nonlinearities, which is the case for e.g., $\tanh(\mathbf{y})$ or $\mathbf{y}^3$, but does not hold for rectification.

While there are more efficient algorithms for nonnegative ICA, such as those based on orthonormal rotations [3], we consider algorithm (5) to be important since it is a "stochastic gradient" or "online" algorithm whereby the parameter matrix $\mathbf{W}$ can be updated on each presentation, rather than requiring a batch of observations. There is also a batch version of (5)

$$\Delta\mathbf{W} = (\mu/p) g_+(\mathbf{Y})(\mathbf{Z} - \mathbf{W}^T g_+(\mathbf{Y}))^T \tag{8}$$

where the columns of $\mathbf{Y}$ and $\mathbf{Z}$ are formed from stacking the individual vectors $\mathbf{y}$ and $\mathbf{z}$ of (5).

In the following sections, we will present the nonnegative PCA algorithm more formally, together with its related ode and give our conjecture on the convergence of this algorithm. We will then analyze the behavior of the algorithm, presenting some analytical results for the multisource (general n) case, with more specific results for $n = 1$ and $n = 2$, including proving convergence of the ode for $n = 1$ and stationarity for $n = 2$. Finally, we will present some illustrative numerical simulations on artificial data and natural images to demonstrate the operation of the algorithm.

II. Algorithm and Conjecture

A. Pre-Whitening

The first stage of our algorithm is a whitening step. In contrast to the pre-whitening step used in many other ICA algorithms, we must be careful not to remove the mean from our data since to do so would lose the information on the nonnegativity of the sources [3]. However, the approach is otherwise similar to standard pre-whitening (see, e.g., [1]).

Given a sequence of observed real data vectors $\mathbf{x}$, let $\mathbf{\Sigma_x} = E\{(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}})^T\}$ be its covariance matrix, where $\bar{\mathbf{x}}$ is the mean of $\mathbf{x}$. Let us form the eigenvector–eigenvalue decomposition $\mathbf{\Sigma_x} = \mathbf{EDE}^T$, where $\mathbf{D} = \mathrm{diag}(d_1,\ldots,d_n)$ is the diagonal matrix containing the (real) eigenvalues of $\mathbf{\Sigma_x}$, and $\mathbf{E} = (\mathbf{e}_1,\ldots,\mathbf{e}_n)$ is the matrix whose columns are the corresponding unit-norm eigenvectors. Then let the whitened data be

$$\mathbf{z} = \mathbf{Vx} = \mathbf{VAs} \tag{9}$$

where $\mathbf{V} = \mathbf{RD}^{-1/2}\mathbf{E}^T$ for any orthonormal matrix $\mathbf{R}$, i.e., such that $\mathbf{R}^T\mathbf{R} = \mathbf{RR}^T = \mathbf{I}$. We often choose $\mathbf{R} = \mathbf{I}$ or $\mathbf{R} = \mathbf{E}$. It is straightforward to verify that $\mathbf{\Sigma_z} = E\{(\mathbf{z}-\bar{\mathbf{z}})(\mathbf{z}-\bar{\mathbf{z}})^T\} = \mathbf{I}$. For unit variance sources, where $\mathbf{\Sigma_s} = E\{(\mathbf{s}-\bar{\mathbf{s}})(\mathbf{s}-\bar{\mathbf{s}})^T\} = \mathbf{I}$ we note that $\mathbf{I} = \mathbf{VA\Sigma_s}(\mathbf{VA})^T = (\mathbf{VA})(\mathbf{VA})^T$ and consequently the $n \times n$ matrix product $(\mathbf{VA})$ is orthonormal.

Note that the covariance matrix $\mathbf{\Sigma_x}$ was calculated with the data mean removed (it is a covariance matrix rather than a correlation matrix), while the mean is not removed during the calculation of the "whitened" data $\mathbf{z}$ from the observed data $\mathbf{x}$. Nevertheless, the covariance matrix $\mathbf{\Sigma_z}$ of $\mathbf{z}$ is a unit matrix, so we refer to this step as a "whitening" process in this article. While zero-mean whitening is convenient for many ICA algorithms, independent random variables need not be zero-mean, and clearly cannot be zero-mean if they are nonnegative (unless they are identically zero). The importance of the whitening step is that the independent sources are an orthogonal rotation of the resulting whitened data $\mathbf{z}$, whether or not the data mean is subtracted during the whitening process.

B. Parameter Update Step

Following pre-whitening, for an online or stochastic gradient algorithm we repeatedly calculate $\mathbf{y} = \mathbf{Wz}$ for each pre-whitened observation vector as it arrives, and apply an update $\mathbf{W} \leftarrow \mathbf{W} + \Delta\mathbf{W}$ using the nonnegative PCA rule (5).

Alternatively, if we have data in the form of a batch of vectors $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_p)$, then we form pre-whitened data $\mathbf{Z} = \mathbf{VX}$. We then repeatedly calculate $\mathbf{Y} = \mathbf{WZ}$, followed by the nonnegative PCA batch rule (8). In both cases, we repeat until the error criterion

$$e_{\mathrm{MSE}} = E\{|\mathbf{z} - \mathbf{W}^T g_+(\mathbf{y})|^2\} \quad \text{(Online)} \tag{10}$$

$$e_{\mathrm{MSE}} = \frac{1}{p}||\mathbf{Z} - \mathbf{W}^T g_+(\mathbf{Y})||_F^2 \quad \text{(Batch)} \tag{11}$$

is sufficiently small, where $||\mathbf{M}||_F^2$ is the square of the Frobenius norm of a matrix $\mathbf{M}$, i.e., the sum of the squares of the elements.

C. Main Conjecture

In general, difference equations such as (5) are very hard to analyze precisely, and it is common to consider instead the behavior of the related ordinary differential equation (ode)

$$d\mathbf{W}/dt = E\{g_+(\mathbf{y})(\mathbf{z} - \mathbf{W}^T g_+(\mathbf{y}))^T\}. \tag{12}$$

For certain reasonable conditions, it can be shown that the difference equation (5) will converge to the asymptotically stable points of (12) (see, e.g., [20] for a discussion of this technique for linear PCA models). We conjecture that (12) converges to find nonnegative sources, and so the nonnegative PCA rule (5) will also tend to find these.

Conjecture 1: Let $\mathbf{s} = (s_1,\ldots,s_n)$ be a vector of well-grounded nonnegative independent sources with bounded nonzero variance and let $\mathbf{z} = \mathbf{VAs}$ where $\mathbf{VA}$ is square and orthonormal, i.e., $(\mathbf{VA})^T\mathbf{VA} = \mathbf{I}$. Then the ode (12) converges to some $\mathbf{W}_\infty$ such that $\mathbf{H} = \mathbf{W}_\infty\mathbf{VA}$ is a nonnegative permutation matrix, so that $\mathbf{y} = \mathbf{Hs}$ is some noninverted permutation of the sources, if and only if the initial value $\mathbf{W}(0)$ is such that $\Pr(y_i > 0) > 0$ for all $1 \le i \le n$.

Remark: The preconditions of this conjecture are satisfied if the sources are independent, and $\mathbf{z}$ is the vector of pre-whitened observations. In fact, we believe that second-order independence ($E\{(s_i - \bar{s}_i)(s_j - \bar{s}_j)\} = 0$ for all $i \neq j$) is likely to be sufficient for convergence of the nonnegative PCA ode (12), although for simplicity of later analysis in this paper we assume the sources are independent.

III. Analysis of the Ode

As mentioned above, one of us previously analyzed the nonlinear PCA ode corresponding to (6) for the case where $g(\cdot)$ is odd and twice differentiable, and $\mathbf{z}$ is zero-mean [18]. Under certain conditions it was shown that the point $\mathbf{W}$ for which $\mathbf{H} = \mathbf{WVA}$ is a suitable diagonal matrix, is asymptotically stable. However, a preliminary analysis suggests that the extension of this approach to the nonnegative PCA rule (5) is not particularly straightforward.

Therefore this paper takes an alternative approach. We shall partition the data space $\mathcal{Z} = \{\mathbf{z}\}$ into different regions for which different subsets of outputs y_i are nonnegative. We will then use known results from the linear PCA subspace algorithm to investigate the behavior due to data arrival in each subset. The overall flow of the ode will be the sum of the flows due to each data region.

A. Linear PCA Subspace Ode

Let us briefly review the behavior of the linear PCA subspace ode [21]. The input vectors $\mathbf{x} = (x_1,\ldots,x_n)$ are used without pre-whitening, so we have $\mathbf{z} = \mathbf{x}$ (i.e., $\mathbf{V} = \mathbf{I}$) with an output $\mathbf{y} = (y_1,\ldots,y_m) = \mathbf{Wx}$ where $\mathbf{W}$ is an $m \times n$ weight matrix with $m \le n$, updated according to the ode

$$d\mathbf{W}/dt = \mathbf{WC}(\mathbf{I} - \mathbf{W}^T\mathbf{W}) \tag{13}$$

where $\mathbf{C} = E\{\mathbf{zz}^T\} = E\{\mathbf{xx}^T\}$ is the correlation matrix of $\mathbf{x}$. We are careful to distinguish the correlation matrix $\mathbf{C}$ from the covariance matrix $\mathbf{\Sigma_x} = E\{(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}})^T\}$, since our data is not zero mean. While many results of the analysis of PCA or PCA subspace algorithms are expressed in terms of a "covariance matrix" of zero-mean data, it is clear from (13) that for nonzero-mean data the behavior of the ode is determined by $\mathbf{C}$ and $\mathbf{W}$ only.

Equation (13) has been studied extensively in recent years. If $\mathbf{C}$ is positive definite such that its mth and $m+1$st eigenvalues are distinct, and $\mathbf{W}(0)$ is finite and full rank, then $\mathbf{W}$ converges to a solution where the rows $\mathbf{w}_i^T$ of $\mathbf{W}$ span the principal subspace, i.e., the subspace spanned by the principal components of $\mathbf{C}$ (see, e.g., [21]–[24]).

Also of interest is the behavior of $\mathbf{W}\mathbf{W}^T$. If $\mathbf{C}$ is positive definite and $\mathbf{W}(0)$ is finite and full rank, then $\mathbf{W}$ remains finite and full rank for all $t \geq 0$, the Lyapunov function

$$J = \|\mathbf{W}\mathbf{W}^T - \mathbf{I}_m\|_F^2 \tag{14}$$

is a nonincreasing function of t except when $\mathbf{W}\mathbf{W}^T = \mathbf{I}$, where $dJ/dt = 0$. Therefore $\mathbf{W}$ converges to a solution of $\mathbf{W}\mathbf{W}^T = \mathbf{I}$, i.e., $\mathbf{W}$ becomes an orthonormal matrix [23], [24]. In fact, J is bounded from above by a curve of the form $J \leq J(0)\exp(-\beta t)$ for some $\beta > 0$ ([23] Proposition 3.1), so given any $\gamma > 0, J(t) < \gamma$ for all $t > t_\gamma = (1/\beta)\log(J(0)/\gamma)$.

B. Multisource Case

While a full analysis of the multi-unit (general n) case has proved to be difficult, there are some results that come immediately. For example, it is straightforward to verify the "only if" direction of Conjecture 1, i.e., that the initial condition on $\mathbf{W}(0)$ is necessary.

Lemma 1: For the system of Conjecture 1, suppose that $\mathbf{W}(0)$ is such that $\Pr(y_{i'} > 0) = 0$ for some i'. Then the system cannot converge to a state where $\mathbf{y}$ is a noninverted permutation of the sources and, hence, $\mathbf{H}$ cannot converge to a nonnegative permutation matrix.

Proof: Let us write $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)^T$ where $\mathbf{w}_i$ is the ith column of $\mathbf{W}^T$ (i.e., that $\mathbf{w}_i^T$ is the ith row of $\mathbf{W}$). Then the ode (12) can be written as n simultaneous odes

$$\dot{\mathbf{w}}_i = E\{g_+ y_i(\mathbf{z} - \mathbf{W}^T g_+(\mathbf{y}))^T\} 1 \leq i \leq n \tag{15}$$

from which it is clear that $\dot{\mathbf{w}}_{i'} = 0$ since $\mathbf{w}_{i'}$ is such that $\Pr(g_+(y_{i'}) \neq 0) = \Pr(y_{i'} > 0) = 0$. Thus $y_{i'} = \mathbf{w}_{i'}^T\mathbf{x}$ will remain unchanged during the operation of the ode. But this is a nonpositive signal, and hence does not (and will never) equal any nonnegative nonzero-variance source s_j. ∎

To make much of the following analysis of the nonnegative PCA ode simpler, we can express the ode (12) in terms of $\mathbf{H} = \mathbf{W}\mathbf{V}\mathbf{A}$ instead of $\mathbf{W}$ [18]. Since $\mathbf{V}\mathbf{A}$ is a constant matrix such that $(\mathbf{V}\mathbf{A})^T(\mathbf{V}\mathbf{A}) = \mathbf{I}$ it is straightforward to show that

$$d\mathbf{H}/dt = E\{g_+(\mathbf{y})(\mathbf{s} - \mathbf{H}^T g_+(\mathbf{y}))^T\} \tag{16}$$

which, together with $\mathbf{y} = \mathbf{H}\mathbf{s}$ in place of $\mathbf{y} = \mathbf{W}\mathbf{z}$, is of exactly the same form as (12) with the substitution $\mathbf{z} \to \mathbf{s}$ and $\mathbf{W} \to \mathbf{H}$. Note that $\mathbf{H}$ in (16) is not constrained to be orthonormal. Thus (as for nonlinear PCA), the behavior of the system is determined purely by the source vectors $\mathbf{s}$ and the overall transform $\mathbf{H}$, a property known as *equivariance* [25]. We can also express the initial conditions on $\mathbf{W}(0)$ used in Conjecture 1 as an equivalent condition on the value of $\mathbf{H}$ in (16).

Lemma 2: For the system of Conjecture 1, $\Pr(y_i > 0) > 0$ for all $1 \leq i \leq n$, if and only if each row vector $\mathbf{h}_i^T$ of $\mathbf{H} = [h_{ij}]$ contains at least one strictly positive element, i.e., that for each i there exists at least one j^+ for which $h_{ij^+} > 0$.

Proof: Intuitively, we can see that due to the well-groundedness of the sources with negative h_{ij}, and the finite variances of the sources with positive h_{ij}, then there will be a nonzero probability that the positives will outweigh the negatives as they sum to give each y_i. To show this more rigorously, let $\mathcal{J} = \{j \mid h_{ij} > 0\}$ be the set of indexes to the strictly positive components of $\mathbf{h}_i$, so $y_i = y_i^{(+)} + y_i^{(-)}$ where $y_i^{(+)} = \sum_{j\in\mathcal{J}} h_{ij}s_j$ and $y_i^{(-)} = \sum_{j\notin\mathcal{J}} h_{ij}s_j$. Now, for each $j \in \mathcal{J}$ we have $\Pr(s_j > \bar{s}_j) > 0$ where $\bar{s}_j$ is the mean of s_j since it has nonzero variance. Thus, for $\alpha = \sum_{j\in\mathcal{J}} h_{ij}\bar{s}_j$

$$\Pr\left(y_i^{(+)} > \alpha\right) > \Pr(h_{ij}s_j > h_{ij}\bar{s}_j \forall j \in \mathcal{J}) \tag{17}$$

$$= \prod_{j\in\mathcal{J}} \Pr(s_j > \bar{s}_j) \tag{18}$$

$$> 0 \tag{19}$$

where we use the fact that the sources are independent. Now, since the sources are well-grounded, for each $j \notin \mathcal{J}$ (for which h_{ij} is negative or zero) we have $\Pr(-h_{ij}s_j < \alpha/(n - |\mathcal{J}|)) = \Pr(s_j < \delta_j) > 0$ using the positive value $\delta_j = -h_{ij}\alpha/(n - |\mathcal{J}|) > 0$. Thus if $\{j \notin \mathcal{J}\} \neq \emptyset$ then

$$\Pr\left(-y_i^{(-)} < \alpha\right) > \prod_{j\notin\mathcal{J}} \Pr(-h_{ij}s_j < \alpha/(n - |\mathcal{J}|)) > 0 \tag{20}$$

and, hence

$$\Pr(y_i > 0) = \Pr\left(y_i^{(+)} > -y_i^{(-)}\right) \tag{21}$$

$$> \Pr\left(y_i^{(+)} > \alpha\right)\Pr\left(\alpha > -y_i^{(-)}\right) \tag{22}$$

$$> 0. \tag{23}$$

If $\{j \notin \mathcal{J}\} = \emptyset$ then all h_{ij} are strictly positive and the inequality holds trivially. For the converse, suppose tentatively that all elements h_{ij} of $\mathbf{h}_i$ are negative or zero. Then $y_i = \sum_j h_{ij}s_j$ must be negative or zero since all the sources are nonnegative and, consequently, $\Pr(y_i > 0) = 0$. Thus, if $\Pr(y_i > 0) > 0$ there must be at least one strictly positive element h_{ij^+} of $\mathbf{h}_i$. ∎

We now show that the convergence points proposed in conjecture are stationary.

Lemma 3: If $\mathbf{H}$ is a nonnegative permutation matrix, then ode (16) is stationary.

Proof: Since both $\mathbf{s}$ and $\mathbf{H}$ are nonnegative, then so is $\mathbf{y}$, so $g_+(\mathbf{y}) = \mathbf{y}$ and ode (16) becomes $d\mathbf{H}/dt = \mathbf{H}\mathbf{C}(\mathbf{I} - \mathbf{H}^T\mathbf{H})$, i.e., the subspace ode (13) with $\mathbf{W} \to \mathbf{H}$ and $\mathbf{C} = E\{\mathbf{s}\mathbf{s}^T\}$. Now if $\mathbf{H}$ is a permutation matrix, it has elements $\mathbf{h}_{ij} = \delta_{jj_i}$ where $j_i \in \{1, \ldots, n\}$ and $j_i \neq j_{i'}$ for all $i \neq i'$. Thus, it follows that $\mathbf{H}^T\mathbf{H} = \mathbf{I}$ so $d\mathbf{H}/dt = 0$. ∎

We call the domain wherein all elements h_{ij} of $\mathbf{H}$ are nonnegative the *linear domain* of (16). In the linear domain, (16) reduces to the linear subspace ode (13), and it is thus relatively straightforward to investigate the behavior in this domain. Clearly, if $\mathbf{H}$ is a nonnegative permutation matrix (Lemma 3) then it is in the linear domain: the following Lemma investigates the behavior of the more general case for $\mathbf{H}$ in the linear domain.

Lemma 4: Let $\mathbf{H}$ be a bounded full rank matrix in the linear domain of (16) and let $\mathbf{C} = E\{\mathbf{s}\mathbf{s}^T\}$ for $\mathbf{s}$ in (16) be positive definite. Then if $\mathbf{H}$ remains in the linear domain it converges to a permutation matrix.

Proof: We note that $\mathbf{H}\mathbf{H}^T = \mathbf{W}\mathbf{V}\mathbf{A}(\mathbf{V}\mathbf{A})^T\mathbf{W}^T = \mathbf{W}\mathbf{W}^T$ so $J = \|\mathbf{W}\mathbf{W}^T - \mathbf{I}_m\|_F^2 = \|\mathbf{H}\mathbf{H}^T - \mathbf{I}_m\|_F^2$, which is

strictly decreasing over time in the linear domain, as we saw in Section III-A. Therefore, if $\mathbf{H}$ remains in the linear domain, it must converge to an orthonormal matrix and, if $\mathbf{H}$ is an orthonormal matrix with all nonnegative elements, it must be a permutation matrix ([2] Lemma 1). ■

Remark: Unit variance of all sources s_i is sufficient to fix $\mathbf{C}$ to be positive definite, since $\mathbf{C} = E\{\mathbf{s}\mathbf{s}^T\} = \mathbf{\Sigma_s} + \bar{\mathbf{s}}\bar{\mathbf{s}}^T = \mathbf{I} + \bar{\mathbf{s}}\bar{\mathbf{s}}^T$.

Thus if $\mathbf{H}$ is full rank and $\mathbf{C}$ is positive definite, the $n!$ permutation matrices are the only stationary points in the linear region, and there are no limit cycles in the linear region. Conjecture 1, if true, would imply these are the *only* stationary points for the nonnegative PCA subspace algorithm, and consequently that there are no stationary points outside of the linear region (except for the "bad" domain of initial $\mathbf{W}(0)$ for which $\Pr(y_i > 0) = 0$ for some i).

The conjecture, if true, would also imply that these points are stable, even against perturbations which take $\mathbf{W}$ outside of the linear region: Lemma 4 does not in itself require $\mathbf{H}$ to be stable if it is a permutation matrix. For any permutation matrix, $n(n-1)$ of its entries are zero, which is right against the edge of the linear region: any perturbation which takes these zero values negative will take the behavior outside of the linear region. We also note that the conjecture implies that there can be no limit cycles outside of the linear region.

We have seen that J in (14) decreases in the linear domain, and we shall see that there are other conditions where J can be shown to be decreasing. However, there are conditions where J does not decrease, as shown by the following lemma, so e.g., it cannot be used as a Lyapunov function or energy function to demonstrate convergence in the general case.

Lemma 5: Let $J = \|\mathbf{I} - \mathbf{H}\mathbf{H}^T\|_F^2$. Then there exists some $\mathbf{H}$ in (16) for which J is zero but $d\mathbf{H}/dt \neq \mathbf{0}$. Furthermore, there exists some $\mathbf{H}'$ for which J is increasing.

Proof: For the case $n = 2$, construct $\mathbf{H} = (\mathbf{h}_1 \ \ \mathbf{h}_2)^T$ where $\mathbf{h}_1 = (h_{11} \ \ h_{12})^T$ with $h_{11} > 0, h_{12} > 0$ and $|\mathbf{h}_1| = 1$, and $\mathbf{h}_2 = (-h_{12}, h_{11})^T$ so that $\mathbf{h}_1^T\mathbf{h}_2 = 0$. It follows that $\mathbf{H}\mathbf{H}^T = \mathbf{H}^T\mathbf{H} = \mathbf{I}$, so $J = 0$. After some manipulation we find $d\mathbf{h}_2^T/dt = 0$ and $d\mathbf{h}_1^T/dt = E_{y_2 \leq 0}\{y_1(\mathbf{s}^T - y_1\mathbf{h}_1^T)\}$, yielding $d/dt(\mathbf{h}_1^T\mathbf{h}_2) = E_{y_2 \leq 0}\{y_1 y_2\} < 0$ since $y_1 \geq 0$, where the inequality is strictly held since $\Pr(y_1 > 0, y_2 < 0) > 0$ due to the independence and well-groundedness of the nonnegative sources s_1, s_2. Therefore $d\mathbf{H}/dt \neq 0$.

Furthermore, let $\mathbf{H}' = (\mathbf{h}_1' \ \ \mathbf{h}_2)^T$ where $\mathbf{h}_1' = (h_{11}(h_{12} - \epsilon))^T$ for some small $\epsilon > 0$ and $\mathbf{h}_2$ is unchanged. This leads eventually to $dJ/dt = -4\epsilon h_{11} E_{y_2 \leq 0}\{y_1 y_2\} + O(\epsilon^2) > 0$, which means that J is not even nonincreasing, so certainly cannot be used as a Lyapunov function (or even a weaker "energy function") here. ■

C. Single-Source Case

The $n = 1$ case is somewhat trivial, but nevertheless worth stating to confirm that the conjecture does not break for this simple case. Here, we have $y = wz = wvx = wvas = hs$ with $va = \pm 1$. Our weight flow for h is given by

$$dh/dt = \begin{cases} h(1-h^2)c_s, & \text{if } h > 0 \\ 0, & \text{otherwise} \end{cases} \tag{24}$$

where $c_s = E\{s^2\}$. This ode clearly converges to $h = 1$ (the trivial "permutation matrix") if $h(0) > 0$, (also given by Lemma 4), and h remains static otherwise. Thus, Conjecture 1 is valid for $n = 1$.

In fact, there is a slightly less trivial version for the case $n = 1$. Suppose that we have some observation $\mathbf{x} = \mathbf{a}s$, for which $\mathbf{\Sigma_x}$ has only one nonzero eigenvalue $|\mathbf{a}|^2$ and corresponding unit eigenvector $\pm\mathbf{a}/|\mathbf{a}|$. In this case, the eigenvalue matrix $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_n) = \mathrm{diag}(|\mathbf{a}|^2, 0, \ldots, 0)$ is noninvertible, so instead of whitening using $\mathbf{D}^{-1/2}$ we could "pseudowhiten" using the square root of pseudo-inverse, $(\mathbf{D}^+)^{1/2} = \mathrm{diag}(|\mathbf{a}|^{-1}, 0, \ldots, 0)$. This gives a pseudowhitening matrix $\mathbf{V} = \mathbf{R}(\mathbf{a}/(|\mathbf{a}|^2), \mathbf{0}, \ldots, \mathbf{0})^T$ for some orthonormal $\mathbf{R}$, leading to $\mathbf{z} = \mathbf{R}(\mathbf{a}^T\mathbf{a}/(|\mathbf{a}|^2), 0, \ldots, 0)^T s$ i.e., $\mathbf{z} = \mathbf{R}(s, 0, \ldots, 0)^T = \mathbf{r}_1 s$, where $\mathbf{r}_1$ is the first column vector of $\mathbf{R}$, and is the only nonzero eigenvector of $\mathbf{\Sigma_z} = E\{(\mathbf{z} - \bar{\mathbf{z}})(\mathbf{z} - \bar{\mathbf{z}})^T\}$.

If we then have $y = \mathbf{w}^T\mathbf{z} = \mathbf{w}^T\mathbf{r}_1 s$, y is either entirely nonnegative ($\Pr(y < 0) = 0$), when $\mathbf{w} \cdot \mathbf{r}_1 \geq 0$ or entirely nonpositive ($\Pr(y > 0) = 0$), when $\mathbf{w} \cdot \mathbf{r}_1 \leq 0$. If nonpositive, the algorithm is stationary. If nonnegative, the algorithm will operate as a linear Oja PCA neuron [26] and, hence, in this particular case converge to $\mathbf{w} = +\mathbf{r}_1$, again leading to $h = 1$.

We suspect that this may be a special case of a more general result. Specifically, we believe that for input data vectors of dimension $m > n$ the nonnegative PCA subspace algorithm will find the subspace containing the pseudowhitened sources for $n > 1$, but we will not consider this further in this particular paper.

D. Two-Source Case: Stationarity

For reasons of space, we shall restrict ourselves to consideration of the stationarity only of the $n = 2$ case in the present paper. We have seen that, in order to analyze the behavior of the nonnegative PCA subspace algorithm as an ICA algorithm, it is sufficient to consider the ode (16) expressed in terms of the source-to-output matrix $\mathbf{H}$, and the sources $\mathbf{s}$, for which $\mathbf{y} = \mathbf{H}\mathbf{s}$. For analysis of convergence, we no longer need to consider what happens to $\mathbf{W}$, $\mathbf{x}$ or $\mathbf{z}$ since these can be reconstructed from $\mathbf{H}$ and $\mathbf{s}$ if required.

However, while this is an ICA algorithm, we will make considerable use of tools from the PCA subspace literature. To assist readers to see more clearly the links to these techniques, we shall take the somewhat unusual step of changing notation for this section. Specifically, we will make the substitutions $\mathbf{s} \to \mathbf{x}$ and $\mathbf{H} \to \mathbf{W}$, leading to a system expressed as

$$\mathbf{y} = \mathbf{W}\mathbf{x} \tag{25}$$

with update ode

$$\dot{\mathbf{W}} = E\{g_+(\mathbf{y})(\mathbf{x} - \mathbf{W}^T g_+(\mathbf{y}))^T\} \tag{26}$$

with well-grounded nonnegative independent sources $\mathbf{x}$. Alternatively, the reader may wish to consider that we have made the assumption, without loss of generality, that $\mathbf{V} = \mathbf{A} = \mathbf{I}$ such that $\mathbf{z} = \mathbf{x} = \mathbf{s}$ and $\mathbf{H} = \mathbf{W}$.

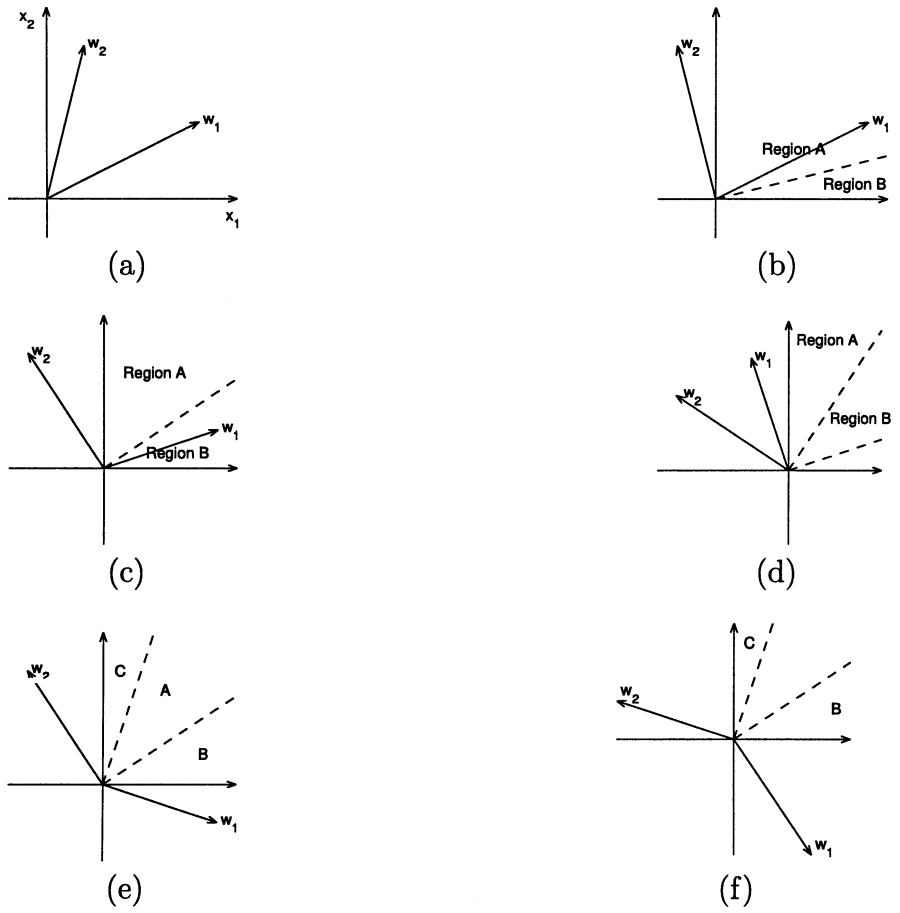

Fig. 1. Some possible arrangements for $\mathbf{W}$, showing the regions $\mathcal{A}, \mathcal{B}$ and $\mathcal{C}$, which are all subsets of the positive quadrant. Region $\mathcal{D}$ is omitted, since it does not lead to any change in $\mathbf{W}$. The $\mathbf{W}$ domains illustrated here are: (a) $\mathcal{W}_{\mathcal{A}}$; (b), (c), and (d) $\mathcal{W}_{\mathcal{AB}}$; (e) $\mathcal{W}_{\mathcal{ABC}}$; and (f) $\mathcal{W}_{\mathcal{BC}}$.

Let us partition the whole data region $\mathcal{X} = \{\mathbf{x}\}$ into four mutually exclusive regions $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}$

$$\mathcal{A} = \{\mathbf{x} \,|\, y_1 \geq 0, y_2 \geq 0\} \tag{27}$$
$$\mathcal{B} = \{\mathbf{x} \,|\, y_1 > 0, y_2 < 0\} \tag{28}$$
$$\mathcal{C} = \{\mathbf{x} \,|\, y_1 < 0, y_2 > 0\} \tag{29}$$
$$\mathcal{D} = \{\mathbf{x} \,|\, y_1 \leq 0, y_2 \leq 0\} - \{\mathbf{x} \,|\, y_1 = y_2 = 0\} \tag{30}$$

where it is straightforward to verify that $\mathcal{A} \cup \mathcal{B} \cup \mathcal{C} \cup \mathcal{D} = \mathcal{X}$. Note that these regions will depend on $\mathbf{W}$ (see Fig. 1 for some possible arrangements).

We can now write the ode (26) as

$$\dot{\mathbf{W}} = \dot{\mathbf{W}}|_{\mathcal{A}} + \dot{\mathbf{W}}|_{\mathcal{B}} + \dot{\mathbf{W}}|_{\mathcal{C}} + \dot{\mathbf{W}}|_{\mathcal{D}} \tag{31}$$

where for any $\mathcal{S} \in \{\mathcal{A}, \mathcal{B}, \mathcal{C}\}$ we have $\dot{\mathbf{W}}|_{\mathcal{S}} = E_{\mathcal{S}}\{g_+(\mathbf{y})(\mathbf{x} - \mathbf{W}^T g_+(\mathbf{y}))^T\}$ with $E_{\mathcal{S}}\{f(\mathbf{x})\} = \int_{\mathcal{S}} f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$ such that $E_{\mathcal{S}}\{1\} = p_{\mathcal{S}} = \Pr(\mathbf{x} \in \mathcal{S})$. Working out the terms in (31) we get

$$\dot{\mathbf{W}}|_{\mathcal{A}} = \mathbf{W}\mathbf{C}_{\mathcal{A}}(\mathbf{I} - \mathbf{W}^T\mathbf{W}) \tag{32}$$
$$\dot{\mathbf{W}}|_{\mathcal{B}} = \begin{bmatrix} \dot{\mathbf{w}}_1^T|_{\mathcal{B}} \\ 0 \end{bmatrix} \quad \dot{\mathbf{w}}_1^T|_{\mathcal{B}} = \mathbf{w}_1^T\mathbf{C}_{\mathcal{B}}\left(\mathbf{I} - \mathbf{w}_1\mathbf{w}_1^T\right) \tag{33}$$
$$\dot{\mathbf{W}}|_{\mathcal{C}} = \begin{bmatrix} 0 \\ \dot{\mathbf{w}}_1^T|_{\mathcal{C}} \end{bmatrix} \quad \dot{\mathbf{w}}_2^T|_{\mathcal{C}} = \mathbf{w}_2^T\mathbf{C}_{\mathcal{C}}\left(\mathbf{I} - \mathbf{w}_2\mathbf{w}_2^T\right) \tag{34}$$
$$\dot{\mathbf{W}}|_{\mathcal{D}} = 0 \tag{35}$$

TABLE I
WEIGHT–SPACE PARTITION FOR STATIONARITY ANALYSIS

$p_{\mathcal{A}}$	$p_{\mathcal{B}}$	$p_{\mathcal{C}}$	Weight Domain
> 0	$= 0$	$= 0$	$\mathcal{W}_{\mathcal{A}}$
> 0	> 0	$= 0$	$\mathcal{W}_{\mathcal{AB}}$
> 0	$= 0$	> 0	$\mathcal{W}_{\mathcal{AC}}$
> 0	> 0	> 0	$\mathcal{W}_{\mathcal{ABC}}$
$= 0$	> 0	> 0	$\mathcal{W}_{\mathcal{BC}}$
$= 0$	$= 0$	> 0	$\mathcal{W}_{\mathcal{C}}$
$= 0$	> 0	$= 0$	$\mathcal{W}_{\mathcal{B}}$
$= 0$	$= 0$	$= 0$	$\mathcal{W}_{\emptyset}$

where $\mathbf{C}_{\mathcal{S}} = E_{\mathcal{S}}\{\mathbf{x}\mathbf{x}^T\}$ for $\mathcal{S} \in \{\mathcal{A}, \mathcal{B}, \mathcal{C}\}$. It immediately follows that (31) does not depend on any $\mathbf{x} \in \mathcal{D}$, so $\dot{\mathbf{W}}|_{\mathcal{D}}$ is omitted in the sequel. We sometimes find it convenient to write (26) in terms of the individual weight vectors as

$$\dot{\mathbf{w}}_1^T = \mathbf{w}_1^T\mathbf{C}_{\mathcal{A}}(\mathbf{I} - \mathbf{W}^T\mathbf{W}) + \mathbf{w}_1^T\mathbf{C}_{\mathcal{B}}\left(\mathbf{I} - \mathbf{w}_1\mathbf{w}_1^T\right) \tag{36}$$
$$\dot{\mathbf{w}}_2^T = \mathbf{w}_2^T\mathbf{C}_{\mathcal{A}}(\mathbf{I} - \mathbf{W}^T\mathbf{W}) + \mathbf{w}_2^T\mathbf{C}_{\mathcal{C}}\left(\mathbf{I} - \mathbf{w}_2\mathbf{w}_2^T\right). \tag{37}$$

We will now investigate the stationarity of (31) and, hence, (26), depending on the various values for $p_{\mathcal{A}}, p_{\mathcal{B}}$, and $p_{\mathcal{C}}$. We partition the parameter space $\mathcal{W} = \{\mathbf{W}\}$ into eight mutually exclusive domains, depending on which of $p_{\mathcal{A}}, p_{\mathcal{B}}$, and $p_{\mathcal{C}}$ are zero (Table I) and noting that $p_{\mathcal{S}} = 0$ implies $\mathbf{C}_{\mathcal{S}} = 0$.

The last three correspond to regions which do not satisfy the initial conditions for Conjecture 1. We therefore wish to show that for any $\mathbf{W} \in \mathcal{W}_{\mathcal{A}} \cup \mathcal{W}_{\mathcal{AB}} \cup \mathcal{W}_{\mathcal{AC}} \cup \mathcal{W}_{\mathcal{ABC}} \cup \mathcal{W}_{\mathcal{AC}}$, the system is stationary if and only if $\mathbf{W}$ is a positive permutation matrix.

Lemma 6: In the system (25) with $n = 2$, and where $\mathbf{x}$ is a two-dimensional (2-D) vector of well-grounded nonnegative unit-variance independent sources, let $\mathbf{W}$ be a bounded full rank 2×2 matrix. If $\mathbf{W} \in \mathcal{W}_{\mathcal{A}}$, then the ode (26) is stationary if and only if $\mathbf{W}$ is a positive permutation matrix.

Proof: The "only if" direction follows immediately from Lemma 4 with an appropriate change of notation. The converse is a special case of Lemma 1 with $n = 2$. ∎

We shall also note the condition for being in $\mathcal{W}_{\mathcal{A}}$, which we will need later.

Lemma 7: For the system of Lemma 6, $\mathbf{W} \in \mathcal{W}_{\mathcal{A}}$ if and only if all the elements of $\mathbf{W}$ are nonnegative.

If any element w_{ij} of $\mathbf{W}$ is negative, there is a nonzero probability that source s_j might outweigh all the others, since they are well-grounded so will sometimes be close to zero, producing a negative y_i. A formal proof along the lines of Lemma 1 can be constructed if desired, but will be omitted here for brevity.

Lemma 8: For the system of Lemma 6, if $\mathbf{W} \in \mathcal{W}_{\mathcal{AB}} \cup \mathcal{W}_{\mathcal{AC}}$, then the ode (26) is nonstationary.

Proof: In $\mathcal{W}_{\mathcal{AB}}$ we have

$$\dot{\mathbf{w}}_1^T = \mathbf{w}_1^T \mathbf{C}_{\mathcal{A}} (\mathbf{I} - \mathbf{W}^T \mathbf{W}) + \mathbf{w}_1^T \mathbf{C}_{\mathcal{B}} \left(\mathbf{I} - \mathbf{w}_1 \mathbf{w}_1^T\right) \quad (38)$$

$$\dot{\mathbf{w}}_2^T = \mathbf{w}_2^T \mathbf{C}_{\mathcal{A}} (\mathbf{I} - \mathbf{W}^T \mathbf{W}). \quad (39)$$

Suppose tentatively that there is a stationary point $\mathbf{W} \in \mathcal{W}_{\mathcal{AB}}$. Let us construct the following function

$$g = \dot{\mathbf{w}}_1^T \mathbf{W}^T \mathbf{W} \mathbf{C}_{\mathcal{A}} \mathbf{w}_2 - \dot{\mathbf{w}}_2^T \left(\mathbf{W}^T \mathbf{W} \mathbf{C}_{\mathcal{A}} + \mathbf{w}_1 \mathbf{w}_1^T \mathbf{C}_{\mathcal{B}}\right) \mathbf{w}_1 \quad (40)$$

which must be zero at any stationary point. Substituting in the relevant expressions for $\dot{\mathbf{w}}_1^T$ and $\dot{\mathbf{w}}_2^T$, and using $\mathbf{W}^T \mathbf{W} = \mathbf{w}_1 \mathbf{w}_1^T + \mathbf{w}_2 \mathbf{w}_2^T$, after some manipulation we get $g = \mathbf{w}_1^T \mathbf{C}_{\mathcal{B}} \mathbf{w}_2 \mathbf{w}_2^T \mathbf{C}_{\mathcal{A}} \mathbf{w}_2 = E_{\mathcal{B}}\{y_1 y_2\} E_{\mathcal{A}}\{y_2^2\} < 0$ since $y_2 \geq 0$ in region $\mathcal{A}$ and $y_1 \geq 0 > y_2$ in region $\mathcal{B}$, and the expectations are nonzero due to the well-groundedness of x_2 and nonzero variance of x_1. Thus, g is nonzero for any $\mathbf{W} \in \mathcal{W}_{\mathcal{AB}}$, so there are no stationary points in $\mathcal{W}_{\mathcal{AB}}$. The result for $\mathcal{W}_{\mathcal{AC}}$ follows immediately by substituting $\mathbf{w}_1 \leftrightarrow \mathbf{w}_2$. ∎

Before we consider $\mathcal{W}_{\mathcal{ABC}}$, let us state a minor result that is intuitively clear from Fig. 1(e) and (f).

Lemma 9: If $p_{\mathcal{B}} > 0$ and $p_{\mathcal{C}} > 0$, i.e., $\mathbf{W} \in \mathcal{W}_{\mathcal{ABC}} \cup \mathcal{W}_{\mathcal{BC}}$, then $\mathbf{w}_1^T \mathbf{w}_2 < 0$.

Proof: Let us construct $\mathbf{X} = [\mathbf{x}_1 \mathbf{x}_2]$, where $\mathbf{x}_1 \in \mathcal{B}$ and $\mathbf{x}_2 \in \mathcal{C}$ are vectors with positive elements. Then $\mathbf{Y}$ is of the form

$$\mathbf{Y} = \mathbf{W}\mathbf{X} = \begin{bmatrix} + & - \\ - & + \end{bmatrix} \quad (41)$$

where $+$ represents a positive number and $-$ a negative number. Therefore

$$\mathbf{W} = \mathbf{Y}\mathbf{X}^{-1} = \frac{1}{\det \mathbf{X}} \mathbf{Y} \mathrm{adj} \mathbf{X} = \frac{1}{\det \mathbf{X}} \begin{bmatrix} + & - \\ - & + \end{bmatrix} \quad (42)$$

so $\mathbf{w}_1^T \mathbf{w}_2 = (\det \mathbf{X})^{-2}[(- \cdot +) + (+ \cdot -)] < 0$. ∎

Lemma 10: For the system of Lemma 6, if $\mathbf{W} \in \mathcal{W}_{\mathcal{ABC}}$, then the ode (26) is nonstationary.

Proof: With $\dot{\mathbf{w}}_1^T$ and $\dot{\mathbf{w}}_2^T$ given by (36) and (37), respectively, let us tentatively suppose that the system is stationary, i.e., that $\dot{\mathbf{w}}_1^T = \dot{\mathbf{w}}_2^T = \mathbf{0}$. Taking the dot product of ode (36) with $\mathbf{w}_1$ and using $\mathbf{W}^T \mathbf{W} = \mathbf{w}_1 \mathbf{w}_1^T + \mathbf{w}_2 \mathbf{w}_2^T$ we get

$$\dot{\mathbf{w}}_1^T \mathbf{w}_1 = c_{11} \left(1 - l_1^2\right) - c_{12} (\mathbf{w}_1 \cdot \mathbf{w}_2) \quad (43)$$

where $c_{11} = \mathbf{w}_1^T (\mathbf{C}_{\mathcal{A}} + \mathbf{C}_{\mathcal{B}}) \mathbf{w}_1 = E\{g_+(y_1)^2\}$, $c_{12} = \mathbf{w}_1^T \mathbf{C}_{\mathcal{A}} \mathbf{w}_2^T = E\{g_+(y_1) g_+(y_2)\}$, and $l_1 = |\mathbf{w}_1|$. From Lemma 9 we have $\mathbf{w}_1 \cdot \mathbf{w}_2 < 0$ for $\mathbf{W} \in \mathcal{W}_{\mathcal{ABC}}$, so we must have $l_1^2 > 1$ for $\mathbf{w}_1$ to be stationary. A similar analysis for $\mathbf{w}_2$ shows that $\dot{\mathbf{w}}_2^T \mathbf{w}_2 = c_{22}(1 - l_2^2) - c_{12}(\mathbf{w}_1 \cdot \mathbf{w}_2)$, where $c_{22} = \mathbf{w}_2^T (\mathbf{C}_{\mathcal{A}} + \mathbf{C}_{\mathcal{C}}) \mathbf{w}_2 = E\{g_+(y_2)^2\}$ and $l_2 = |\mathbf{w}_2|$ and, hence, demonstrates that $l_2^2 > 1$ for $\mathbf{w}_2$ to be stationary.

Now, let us consider the behavior of the nonorthonormality measure $J = (1/2)\|I - \mathbf{W}\mathbf{W}^T\|_{\mathrm{F}}^2$ in the various flow regions $\mathcal{A}, \mathcal{B}$, and $\mathcal{C}$. If J decreases under the flow due to each region, then it must decrease under the composite flow due to the sum of the effects of these regions. Let us write $dJ/dt = dJ/dt|_{\mathcal{A}} + dJ/dt|_{\mathcal{B}} + dJ/dt|_{\mathcal{C}}$ where $dJ/dt|_{\mathcal{S}} = \langle \nabla_{\mathbf{W}} J \dot{\mathbf{W}}|_{\mathcal{S}} \rangle$ for $\mathcal{S} \in \{\mathcal{A}, \mathcal{B}, \mathcal{C}\}$. Since region $\mathcal{A}$ corresponds to the linear case, we already know that $dJ/dt|_{\mathcal{A}} \leq 0$ with equality if and only if $\mathbf{W}$ is orthonormal, i.e., $\mathbf{W}^T \mathbf{W} = \mathbf{I}$.

Considering region $\mathcal{B}$, we find

$$dJ/dt|_{\mathcal{B}} = -E_{\mathcal{B}}\left\{y_1^2\right\} \left(1 - \mathbf{w}_1^T \mathbf{w}_1\right)^2 + 2\left[E_{\mathcal{B}}\left\{y_1 \mathbf{x}^T \left(\mathbf{I} - \mathbf{w}_1 \mathbf{w}_1^T\right) \mathbf{w}_2\right\}\right] \left(\mathbf{w}_2^T \mathbf{w}_1\right) \quad (44)$$

of which we can immediately see that the first term is nonpositive, and zero if and only if $|\mathbf{w}_1| = 1$. We know that $y_1 > 0$ in region $\mathcal{B}$, and from Lemma 9 we know that $\mathbf{w}_2^T \mathbf{w}_1 < 0$ for $\mathbf{W} \in \mathcal{W}_{\mathcal{ABC}}$. Therefore, if we could show that $\mathbf{x}^T (\mathbf{I} - \mathbf{w}_1 \mathbf{w}_1^T) \mathbf{w}_2 > 0$ for all $\mathbf{x} \in \mathcal{B}$, that would suffice to show that the second term was negative and, hence, that $dJ/dt|_{\mathcal{B}} < 0$.

From the calculation in Lemma 9 we know that $\mathbf{W}$ takes the form

$$\mathbf{W} = \frac{1}{\det \mathbf{X}} \begin{bmatrix} + & - \\ - & + \end{bmatrix} \quad (45)$$

for an appropriate matrix $\mathbf{X}$, so assuming without loss of generality that $\det \mathbf{X} > 0$, we have $w_{11} > 0 > w_{12}$. Let us introduce the orthonormal coordinate transformation matrix $\mathbf{Q} = (1/|\mathbf{w}_1|)(-\mathbf{w}_1 \quad \mathbf{w}_{1^\perp})^T$ where $\mathbf{w}_{1^\perp}^T = (-w_{12} \quad w_{11})$ is the vector $\mathbf{w}_1$ rotated by $\pi/2$ into the positive quadrant. It is a straightforward calculation to verify that $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}$. Considering some $\mathbf{x}' \in \mathcal{A} \neq \emptyset$, we can show that $\mathbf{w}_{1^\perp}^T \mathbf{w}_2 > 0$ so both elements of $\mathbf{Q}\mathbf{w}_2$ are positive.

Now, under our tentative assumption of stationarity, we must have $l_1^2 > 1$. Therefore for any $\mathbf{x} \in \mathcal{B}$, all of which satisfy $\mathbf{w}_1^T \mathbf{x} > 0$, we have $\mathbf{w}_1^T (\mathbf{I} - \mathbf{w}_1 \mathbf{w}_1^T) \mathbf{x} = (1 - l_1^2)(\mathbf{w}_1^T \mathbf{x}) < 0$, and we also have $\mathbf{w}_{1^\perp}^T (\mathbf{I} - \mathbf{w}_1 \mathbf{w}_1^T) \mathbf{x} = \mathbf{w}_{1^\perp}^T \mathbf{x} > 0$ due to

the positive-only elements in both. Therefore, both elements of $\mathbf{Q}(\mathbf{I}-\mathbf{w}_1\mathbf{w}_1^T)\mathbf{x}$ are positive, from which we can conclude

$$\mathbf{w}_2 \cdot \left(\left(\mathbf{I}-\mathbf{w}_1\mathbf{w}_1^T\right)\mathbf{x}\right) = (\mathbf{Q}\mathbf{w}_2)\cdot\left(\mathbf{Q}\left(\mathbf{I}-\mathbf{w}_1\mathbf{w}_1^T\right)\mathbf{x}\right) > 0 \tag{46}$$

and, hence, $dJ/dt|_{\mathcal{B}} < 0$. A similar argument shows that $dJ/dt|_{\mathcal{C}} < 0$ and, consequently, that $dJ/dt = dJ/dt|_{\mathcal{A}} + dJ/dt|_{\mathcal{B}} + dJ/dt|_{\mathcal{C}} < 0$. Since J is a function of $\mathbf{W}$, $\mathbf{W}$ cannot be stationary. ■

Finally, let us consider the remaining weight domain $\mathcal{W}_{\mathcal{BC}}$.

Lemma 11: For the system of Lemma 6, if $\mathbf{W} \in \mathcal{W}_{\mathcal{BC}}$, then the ode (26) is nonstationary.

Proof: Consider $\mathbf{w}_1$, which will only be affected by $\mathbf{x} \in \mathcal{B}$. We know from the calculation in Lemma 9 that the two elements of $\mathbf{w}_1$ must have opposite signs. Without loss of generality, let us assume that $w_{12} < 0$, then we calculate $dw_{12}/dt = E_{\mathcal{B}}\{y_1(x_2 - y_1 w_{12})\} > 0$ since $w_{12} < 0$ but all other quantities are positive. Therefore, $\mathbf{W}$ is not stationary in this domain. (A similar result also holds for w_{21}.) ■

We notice that this proof for Lemma 11 also demonstrates that there are no limit cycles in the domain $\mathcal{W}_{\mathcal{BC}}$ since we have a monotonically increasing quantity (w_{12}) in this domain. We are finally in a position to state the main stationarity theorem for this $n = 2$ system.

Theorem 1: Let $\mathbf{W}$ in (25) be a bounded full rank 2×2 matrix, where each row $\mathbf{w}_i^T$ of $\mathbf{W}$ contains at least one strictly positive element, and $\mathbf{x}$ is a 2-D vector of well-grounded nonnegative unit-variance independent sources. Then ode (26) is stationary if and only if $\mathbf{W}$ is a positive permutation matrix.

Proof: The "if" direction is simply a special case of Lemma 3. For the converse, from Lemma 2, we know that the precondition on $\mathbf{W}$ is equivalent to requiring $\mathbf{W} \in \mathcal{W}_{\mathcal{A}} \cup \mathcal{W}_{\mathcal{AB}} \cup \mathcal{W}_{\mathcal{AC}} \cup \mathcal{W}_{\mathcal{ABC}} \cup \mathcal{W}_{\mathcal{AC}}$. From Lemma 7, we know that if $\mathbf{W}$ is positive, it must be in $\mathcal{W}_{\mathcal{A}}$. Therefore, we must show: (a) if $\mathbf{W} \in \mathcal{W}_{\mathcal{A}}$, then ode (26) is stationary if and only if $\mathbf{W}$ is a positive permutation matrix and (b) if $\mathbf{W} \in \mathcal{W}_{\mathcal{AB}} \cup \mathcal{W}_{\mathcal{AC}} \cup \mathcal{W}_{\mathcal{ABC}} \cup \mathcal{W}_{\mathcal{AC}}$, then the system cannot be stationary. Lemma 6 gives us (a) and Lemmas 8, 10, and 11 give us (b). ■

IV. Numerical Simulations

In this section, we present some numerical simulations, run in Matlab, to demonstrate the behavior of the algorithm. The forward matrix $\mathbf{W}$ is initialized to the identity matrix, ensuring initial orthogonality of $\mathbf{W}$ and hence of $\mathbf{H} = \mathbf{WVA}$. For these practical simulations, we augment the basic algorithm with the following control mechanisms, to avoid the "bad" region and to automate the choice of learning rate.

1) At each iteration, if there is any output k with $y_{ik} \le 0$ for all patterns i, then the kth row $\mathbf{w}_k^T$ of $\mathbf{W}$ is inverted to remove us from this undesired weight domain. We have only observed this to be needed at the start, or if the algorithm was about to become unstable due to a large update factor.
2) The update factor μ is automatically adapted for best decrease in mean squared error. We found that a test every 10% increase in iterations, with scaling of 2, 1, or 0.5 gave reasonable performance.
3) A test for instability is performed every iteration: if more than a small increase in MSE is found, the update factor is immediately scaled down by 0.5 and retried.

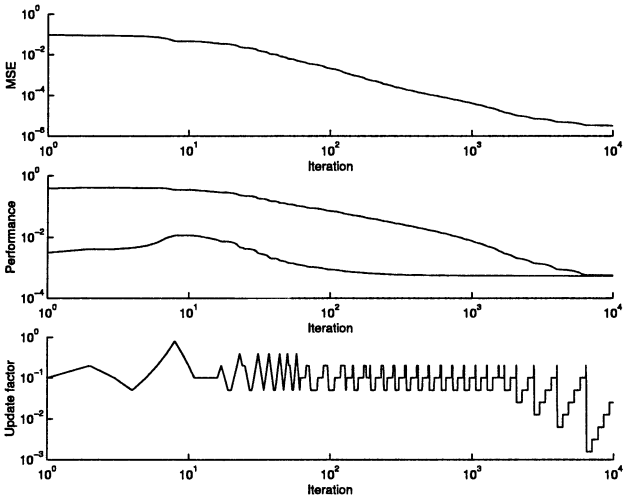

Fig. 2. Simulation on uniformly distributed artificial data for $n = 3$. The top graph shows e_{MSE}, the middle graph showing e_{Perm} (upper curve) and e_{Orth} (lower curve), with the bottom graph showing the update factor μ used.

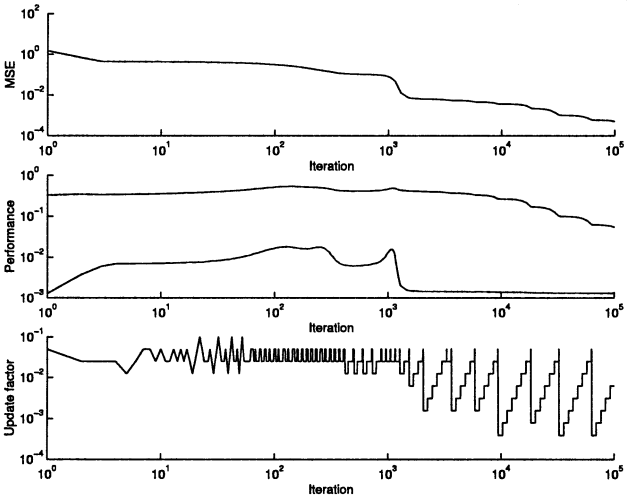

Fig. 3. Simulation on uniformly distributed artificial data for $n = 10$, with the same layout as Fig. 2.

We used the batch update method (8) here, which should have a behavior close to the ode (12). Apart from the adaptive learning rate μ, we have chosen not to use any other acceleration methods, so that we show the behavior of the basic algorithm as closely as possible.

A. Artificial Data

We generated p source vectors $\mathbf{s}_p$ each with n elements randomly using the Matlab 4 uniform generator "rand," scaling the resulting values to lie in the interval $[0, \sqrt{12}]$ to ensure unit variance sources. The $n \times n$ mixing matrix $\mathbf{A}$ was generated randomly from the normally distributed generator "randn," which

TABLE II
PARAMETER MATRIX $\mathbf{H} = \mathbf{WVA}$ FOR $n = 10$ AFTER 10^5 ITERATIONS

$$\mathbf{H} = \begin{pmatrix} 0.043 & -0.047 & -0.110 & -0.125 & -0.108 & 0.472 & -0.003 & \mathbf{0.854} & -0.020 & -0.050 \\ \mathbf{0.641} & 0.009 & -0.069 & -0.042 & -0.038 & \mathbf{0.664} & 0.037 & -0.414 & -0.006 & 0.046 \\ 0.021 & \mathbf{0.972} & 0.042 & 0.024 & 0.000 & 0.054 & -0.035 & 0.059 & -0.029 & -0.012 \\ -0.124 & -0.009 & 0.010 & 0.013 & \mathbf{0.995} & 0.188 & -0.004 & 0.055 & 0.027 & -0.031 \\ \mathbf{0.761} & -0.059 & 0.091 & 0.099 & 0.245 & \mathbf{-0.524} & -0.004 & 0.307 & 0.032 & -0.006 \\ -0.011 & 0.007 & 0.007 & -0.003 & -0.052 & -0.015 & 0.009 & 0.030 & \mathbf{0.993} & 0.026 \\ -0.041 & -0.010 & -0.014 & \mathbf{1.000} & 0.019 & 0.171 & -0.024 & 0.088 & -0.016 & 0.020 \\ -0.044 & 0.005 & 0.042 & -0.096 & 0.036 & -0.009 & -0.004 & 0.056 & 0.012 & \mathbf{0.991} \\ -0.106 & 0.035 & \mathbf{1.002} & 0.056 & -0.074 & 0.130 & -0.037 & 0.027 & -0.019 & -0.040 \\ 0.024 & -0.009 & -0.031 & 0.030 & 0.025 & -0.028 & \mathbf{0.992} & -0.017 & -0.028 & 0.007 \end{pmatrix} \tag{53}$$

has mean zero and variance one. Thus the sources are nonnegative, while the mixing matrix is of mixed sign.

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:

$$e_{\mathrm{MSE}} = \frac{1}{np}\sum_{k=1}^{p} |\mathbf{z}_k - \mathbf{W}^T g_+\{\mathbf{y}_k\}|^2 \tag{47}$$

$$e_{\mathrm{Orth}} = \frac{1}{n^2}\|\mathbf{I} - (\mathbf{WVA})^T\mathbf{WVA}\|_F^2 \tag{48}$$

$$e_{\mathrm{Perm}} = \frac{1}{n^2}\|\mathbf{I} - \mathrm{abs}(\mathbf{WVA})^T\mathrm{abs}(\mathbf{WVA})\|_F^2 \tag{49}$$

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

Fig. 2 shows a typical operation of the algorithm for $n = 3$ sources with $p = 1000$ samples.

After 10^4 iterations (203 s of CPU time on an 850 MHz Pentium 3), the source-to-output matrix $\mathbf{H} = \mathbf{WVA}$ was

$$\mathbf{H} = \begin{pmatrix} -0.0021 & -0.0164 & \mathbf{0.9997} \\ -0.0106 & \mathbf{1.0300} & -0.0027 \\ \mathbf{0.9995} & 0.0022 & -0.0103 \end{pmatrix} \tag{50}$$

with $e_{\mathrm{MSE}} = 3.18 \times 10^{-6}$, $e_{\mathrm{Orth}} = 5.47 \times 10^{-4}$, and $e_{\mathrm{Perm}} = 5.69 \times 10^{-4}$.

Fig. 3 shows a typical operation of the algorithm for $n = 10$ sources with $p = 1000$ samples.

After 10^5 iterations (1.5 h/5750 s of CPU time on an 850-MHz Pentium 3), the source-to-output matrix $\mathbf{H} = \mathbf{WVA}$ was as given in Table II with $e_{\mathrm{MSE}} = 5.02 \times 10^{-4}$, $e_{\mathrm{Orth}} = 1.32 \times 10^{-3}$, and $e_{\mathrm{Perm}} = 0.0553$. From $\mathbf{H}$, we can see that sources (columns) 1 and 6 both have a significant effect on each of the outputs (rows) 5 and 2, although we would expect further learning to remove this. The learning rate variations after 10^3 iterations in Fig. 3 indicate that our current combination of adaptive learning rate and instability avoidance is leading to oscillation of μ and could be improved.

These simulations seem to indicate that the e_{MSE} decreases during the operation of the algorithm. While both e_{Orth} and e_{Perm} are brought close to zero as the algorithm proceeds, it is clear that both of these quantities may increase as well as decrease, as we showed in Lemma 5. We have not observed a situation where the algorithm does not tend to find a permutation matrix, if left long enough (recall that we avoid the undesired $\Pr(y_i > 0) = 0$ domain by flipping $\mathbf{w}_i$ if necessary).

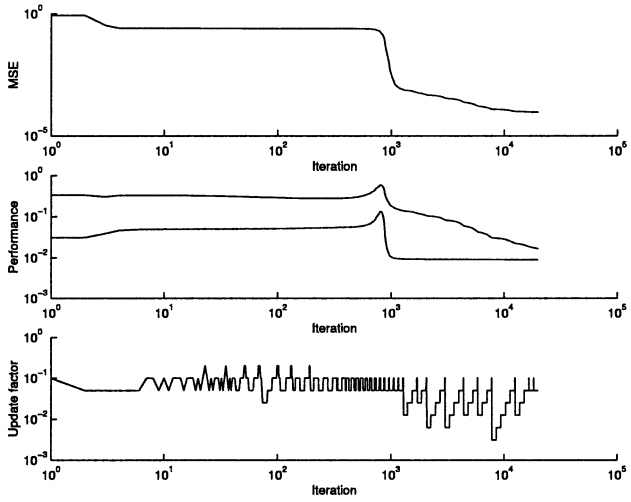

Fig. 4. Simulation results on image data.

B. Natural Image Data

The algorithm was applied to an image unmixing task. For this task, four image patches of size 252×252 were selected from a set of images of natural scenes [27], and downsampled by a factor of 4 in both directions to yield 63×63 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* as required by the algorithm, and the images were scaled to ensure they were all unit variance. After scaling, the source covariance matrix was found to be

$$\overline{\mathbf{ss}^T} - \overline{\mathbf{s}}\,\overline{\mathbf{s}}^T = \begin{pmatrix} 1.000 & 0.074 & -0.003 & 0.050 \\ 0.074 & 1.000 & -0.071 & 0.160 \\ -0.003 & -0.071 & 1.000 & 0.130 \\ 0.050 & 0.160 & 0.130 & 1.000 \end{pmatrix} \tag{51}$$

which showed an acceptably small covariance between the images: as with any ICA method based on pre-whitening, any covariance between sources would prevent accurate identification of the sources. A mixing matrix $\mathbf{A}$ was generated randomly and used to construct $\mathbf{X}$ using the same method as for the uniform data in the previous section.

Fig. 4 shows the performance of learning over 2×10^4 steps, with Fig. 5 showing the original, mixed and separated images and their histograms.

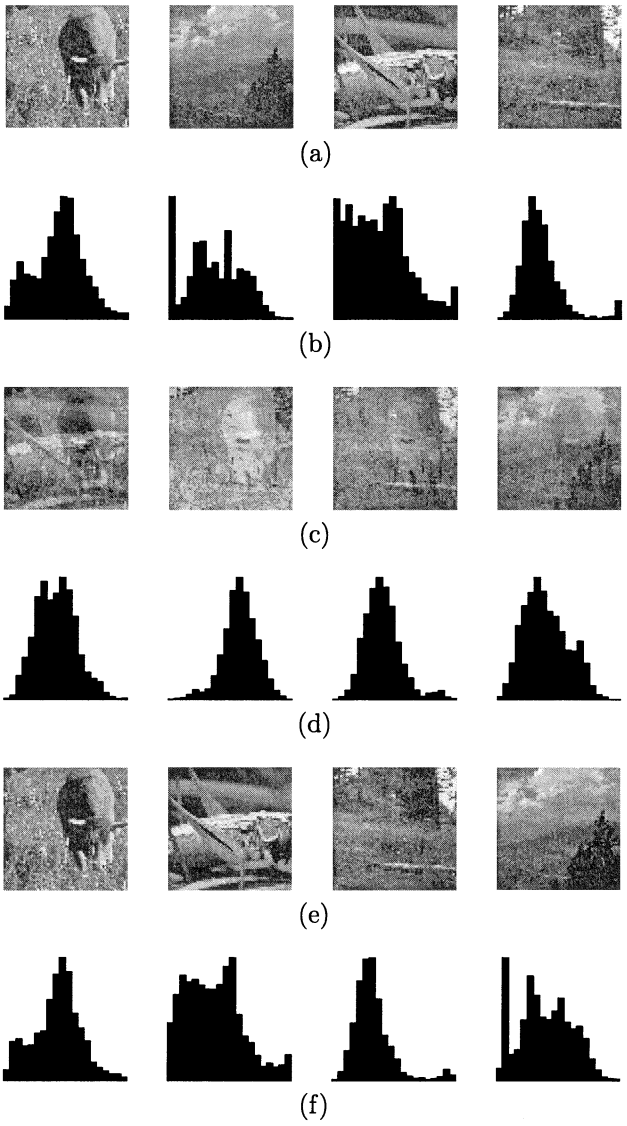

Fig. 5. Images and histograms for the image separation task showing: (a) the original source images and (b) their histograms; (c) and (d) the mixed images and their histograms; and (e) and (f) the separated images and their histograms.

After 2×10^4 iterations (1694 s/28 min of CPU time on an 850 MHz Pentium 3), the source-to-output matrix $\mathbf{H} = \mathbf{WVA}$ was

$$\mathbf{H} = \begin{pmatrix} \mathbf{0.997} & -0.007 & -0.106 & 0.049 \\ -0.126 & 0.042 & -0.078 & \mathbf{1.009} \\ -0.003 & \mathbf{1.007} & -0.102 & 0.046 \\ 0.072 & -0.042 & \mathbf{1.018} & -0.086 \end{pmatrix} \tag{52}$$

with $e_{\mathrm{MSE}} = 9.30 \times 10^{-5}$, $e_{\mathrm{Orth}} = 9.02 \times 10^{-3}$, and $e_{\mathrm{Perm}} = 1.68 \times 10^{-2}$. The algorithm is able to separate the images reasonably well: in fact, good initial separation is already achieved by iteration 10^3, although the algorithm is slow to remove remaining crosstalk between the images.

V. Discussion

The "nonnegative PCA" algorithm (5) considered in this paper is not particularly efficient, and seems to become very slow as n increases. We have concentrated on the batch method here, for closer compatibility with the ode form: the online version is normally faster. Least-squares methods [28] have been used elsewhere to speed up the closely related nonlinear PCA rule (6), and it is possible that this approach could be used here. There are also alternative and more efficient batch algorithms for nonnegative ICA which search over a constrained set of $\mathbf{W}$ [3] and it is also possible to use any standard ICA algorithm followed by inversion of nonpositive sources [29].

The image separation task was constructed in order to visually demonstrate the operation of the algorithm. In general, image separation tasks may be more difficult for this nonnegative PCA algorithm, due to either nongrounded sources, or covariance between the sources. Nongrounded sources could allow the mean squared error to be minimized, while leaving uncertainty or "slack" in the axis rotation to be performed [2]. One way to overcome this might be to attempt to estimate the lower bound of each source as an algorithm progresses, and subtract from each $\mathbf{y}_i$ before rectification. Covariance can also be a significant problem, particular for standard images of faces: these tend to be one color in the center with a contrasting color around the outside and, hence, exhibit significant covariance with each other. It is possible that we could adapt the use of innovations [30] to remove such correlations, while retaining the use of the nonnegativity information.

VI. Conclusion

We have considered a "nonnegative PCA" algorithm for nonnegative independent component analysis, based on the nonlinear PCA algorithms but with a rectification nonlinearity. We conjecture that this algorithm converges to find the nonnegative sources, provided certain initial conditions are met.

We have given some analytical results that are consistent with this conjecture, including stability in the linear region, convergence of the single-source case, and stationarity in the two-source case. We performed numerical simulations on both artificial data and mixtures of natural images, which also indicate that the nonnegative PCA algorithm performs as we expect.

Acknowledgment

One of the authors, M. D. Plumbley would particularly like to thank the Laboratory of Computer and Information Science at HUT for much help and support during the visit, as well as various members of the Neural Networks Research Centre for many interesting and stimulating discussions. The images used in Section IV-B were kindly supplied by P. Hoyer as part of his "imageica" package and are available at http://www.cis.hut.fi/phoyer/NCimages.html. The photographs from which these are derived are © J. Sinkkonen, 1994–1997. Some concepts used in the numerical simulations are also based on the "imageica" package.

References

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[2] M. D. Plumbley, "Conditions for nonnegative independent component analysis," *IEEE Signal Processing Lett.*, vol. 9, pp. 177–180, June 2002.

[3] ——, "Algorithms for nonnegative independent component analysis," *IEEE Trans. Neural Networks*, vol. 14, pp. 534–543, May 2003.
[4] L. Parra, C. Spence, P. Sajda, A. Ziehe, and K.-R. Müller, "Unmixing hyperspectral data," in *Proc. Advances in Neural Information Processing Systems 12 (NIPS'99)*. Denver, CO, 2000, pp. 942–948.
[5] J. S. Lee, D. D. Lee, S. Choi, and D. S. Lee, "Application of nonnegative matrix factorization to dynamic positron emission tomography," in *Proc. Int. Conf. Independent Component Analysis and Signal Separation (ICA'01)*, T.-W. Lee, T.-P. Jung, S. Makeig, and T. J. Sejnowski, Eds., San Diego, CA, Dec. 2001, pp. 629–632.
[6] S. Tsuge, M. Shishibori, S. Kuroiwa, and K. Kita, "Dimensionality reduction using nonnegative matrix factorization for information retrieval," in *Proc. IEEE Int. Conf. Systems, Man, and Cybernetics*, vol. 2, Tucson, AZ, Oct. 2001, pp. 960–965.
[7] R. C. Henry, "Multivariate receptor models-current practice and future trends," *Chemo-metrics Intell. Lab. Syst.*, vol. 60, no. 1–2, pp. 43–48, Jan. 2002.
[8] L. Xu, "Least mean square error reconstruction principle for self-organizing neural-nets," *Neural Networks*, vol. 6, no. 5, pp. 627–648, 1993.
[9] G. F. Harpur, "Low Entropy Coding with Unsupervised Neural Networks," Ph.D. dissertation, Dept. Engineering, University of Cambridge, Cambridge, U.K., Feb. 1997.
[10] D. Charles and C. Fyfe, "Modeling multiple-cause structure using rectification constraints," *Network: Computat. Neural Syst.*, vol. 9, pp. 167–182, 1998.
[11] C. Fyfe, "Positive weights in interneurons," in *Neural Computing: Research and Applications II. Proc. 3rd Irish Neural Networks Conf. Belfast, Northern Ireland*, G. Orchard, Ed. Belfast, NI: Irish Neural Networks Association, Sept. 1993, pp. 47–58.
[12] ——, "A neural net for PCA and beyond," *Neural Processing Lett.*, vol. 6, pp. 33–41, 1997.
[13] P. Paatero and U. Tapper, "Positive matrix factorization: A nonnegative factor model with optimal utilization of error estimates of data values," *Environmetr.*, vol. 5, pp. 111–126, 1994.
[14] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 21, 1999.
[15] K. I. Diamantaras and S. Y. Kung, *Principal Component Neural Networks: Theory and Applications, Adaptive and Learning Systems for Signal Processing, Communications, and Control*. New York: Wiley, 1996.
[16] E. Oja, "Nonlinear PCA criterion and maximum likelihood in independent component analysis," in *Proc. Int. Workshop Independent Component Analysis and Signal Separation (1CA'99)*, Aussois, France, 1999, pp. 143–148.
[17] E. Oja, H. Ogawa, and J. Wangviwattana *et al.*, "Learning in nonlinear constrained Hebbian networks," in *Artificial Neural Networks*, T. Kohonen *et al.*, Eds. Amsterdam, The Netherlands: North-Holland, 1991, pp. 199–202.
[18] E. Oja, "The nonlinear PCA learning rule in independent component analysis," *Neurocomputing*, vol. 17, no. 1, pp. 25–45, 1997.
[19] J. Karhunen and J. Joutsensalo, "Representation and separation of signals using nonlinear PCA type learning," *Neural Networks*, vol. 7, no. 1, pp. 113–127, 1994.
[20] E. Oja and J. Karhunen, "On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix," *J. Mathematical Anal. Applicat.*, vol. 106, pp. 69–84, 1985.
[21] E. Oja, "Neural networks, principal components, and subspaces," *Int. J. Neural Syst.*, vol. 1, pp. 61–68, 1989.
[22] K. Hornik and C.-M. Kuan, "Convergence analysis of local feature extraction algorithms," *Neural Networks*, vol. 5, no. 2, pp. 229–240, 1992.
[23] W.-Y. Yan, U. Helmke, and J. B. Moore, "Global analysis of Oja's flow for neural networks," *IEEE Trans. Neural Networks*, vol. 5, pp. 674–683, 1994.
[24] M. D. Plumbley, "Lyapunov functions for convergence of principal component algorithms," *Neural Networks*, vol. 8, pp. 11–23, 1995.
[25] J.-F. Cardoso and B. H. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Processing*, vol. 44, pp. 3017–3030, 1996.
[26] E. Oja, "A simplified neuron model as a principal component analyzer," *J. Mathematical Biol.*, vol. 15, pp. 267–273, 1982.
[27] A. Hyvärinen and P. Hoyer, "Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces," *Neural Computat.*, vol. 12, no. 7, pp. 1705–1720, 2000.
[28] J. Karhunen and P. Pajunen, "Blind source separation using least-squares type adaptive algorithms," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'97)*, vol. 4, Munich, Germany, 1997, pp. 3361–3364.
[29] A. Cichocki and P. Georgiev, "Blind source separation algorithms with matrix constraints," *IEICE Trans. Fundamentals Electron., Communicat., Comput. Sci.*, vol. E86-A, no. 3, pp. 522–531, Mar. 2003.
[30] A. Hyvärinen, "Independent component analysis for time-dependent stochastic processes," in *Proc. Int. Conf. Artificial Neural Networks (ICANN'98)*, Sweden, 1998, pp. 541–546.

Mark D. Plumbley (S'88–M'90) received the B.A. (later M.A.) degree in electrical sciences from Churchill College, University of Cambridge, U.K., in 1984, respectively, the Graduate Diploma degree in digital systems design from Brunel University, Uxbridge, U.K., in 1986, and the Ph.D. degree in information theory and neural networks from the Department of Engineering, University of Cambridge, U.K., in 1991.

He was with Thorn-EMI Central Research Laboratories in 1986 and began research on neural networks in 1987, as Research Student with the Department of Engineering, University of Cambridge, U.K. He continued there as a Research Associate using genetic algorithms to modify neural networks. He joined the Centre for Neural Networks, Kings College, London, U.K., in 1991, moving to the Department of Electronic Engineering in 1995. In 2002, he joined the new DSP and Multimedia Group, Queen Mary University, London, U.K. He is currently working on independent component analysis (ICA), with a particular interest in applications to the analysis and separation of audio and music signals.

Erkki Oja (S'76–M'77–SM'90–F'00) received the Dr.Sc. degree from Helsinki University of Technology, Helsinki, Finland, in 1977.

He has been a Research Associate with Brown University, Providence, RI, and a Visiting Professor at Tokyo Institute of Technology, Tokyo, Japan. He is Director of the Neural Networks Research Centre and Professor of Computer Science at the Laboratory of Computer and Information Science, Helsinki University of Technology. He is the author or coauthor of more than 240 articles and book chapters on pattern recognition, computer vision, and neural computing, as well as three books: *Subspace Methods of Pattern Recognition* (New York: RSP/Wiley, 1983— transl. Chinese, Japanese), *Kohonen Maps* (Amsterdam, The Netherlands: Elsevier, 1999), and *Independent Component Analysis* (New York: Wiley, 2001). His research interests are in the study of principal components, independent components, self-organization, statistical pattern recognition, and applying artificial neural networks to computer vision and signal processing.

Prof. Oja is a member of the editorial boards of several journals and has been on the program committees of several recent conferences, including ICANN, IJCNN, and ICONIP. He is a member of the Finnish Academy of Sciences, Founding Fellow of the International Association of Pattern Recognition (IAPR), and President of the European Neural Network Society (ENNS).