Nonholonomic Orthogonal Learning Algorithms for Blind Source Separation

Shun-ichi Amari
Tian-Ping Chen
Andrzej Cichocki
RIKEN Brain Science Institute, Brain-Style Information Systems Group, Japan

Independent component analysis or blind source separation extracts independent signals from their linear mixtures without assuming prior knowledge of their mixing coefficients. It is known that the independent signals in the observed mixtures can be successfully extracted except for their order and scales. In order to resolve the indeterminacy of scales, most learning algorithms impose some constraints on the magnitudes of the recovered signals. However, when the source signals are nonstationary and their average magnitudes change rapidly, the constraints force a rapid change in the magnitude of the separating matrix. This is the case with most applications (e.g., speech sounds, electroencephalogram signals). It is known that this causes numerical instability in some cases. In order to resolve this difficulty, this article introduces new nonholonomic constraints in the learning algorithm. This is motivated by the geometrical consideration that the directions of change in the separating matrix should be orthogonal to the equivalence class of separating matrices due to the scaling indeterminacy. These constraints are proved to be nonholonomic, so that the proposed algorithm is able to adapt to rapid or intermittent changes in the magnitudes of the source signals. The proposed algorithm works well even when the number of the sources is overestimated, whereas the existent algorithms do not (assuming the sensor noise is negligibly small), because they amplify the null components not included in the sources. Computer simulations confirm this desirable property.

1 Introduction

Blind source separation is the problem of recovering a number of independent random or deterministic signals when only their linear mixtures are available. Here, “blind” implies that the mixing coefficients, the original signals, and their probability distributions are not known a priori. The problem is also called independent component analysis (ICA) or blind extraction of source signals. The problem has become increasingly important because of many potential applications, such as those in speech recognition and env-

Neural Computation 12, 1463–1484 (2000) © 2000 Massachusetts Institute of Technology
The source signals can be identified only up to a certain ambiguity, that is, up to a permutation and arbitrary scaling of original signals. In order to resolve the scaling indeterminacy, most learning algorithms developed so far impose some constraints on the power or amplitude of the separated signals (Jutten & Hérault, 1991; Hérault & Jutten, 1986; Bell & Sejnowski, 1995; Common, 1994; Douglas, Cichocki, & Amari, 1997; Sorouchyari, 1991; Chen & Chen, 1994; Cichocki, Unbehauen, & Rummert, 1994; Cichocki, Unbehauen, Moszczynski, & Rummert, 1994; Cichocki, Amari, Adachi, & Kasprzak, 1996; Cichocki, Bogner, Moszczynski, & Pope, 1997; Cardoso & Laheld, 1996; Girolami & Fyfe, 1997a, 1997b; Amari, Chen, & Cichocki, 1997; Amari, Cichocki, & Yang, 1995, 1996; Amari & Cardoso, 1997; Amari, 1967; Amari & Cichocki, 1998).

When the source signals are stationary and their average magnitudes do not change over time, the constraints work well to extract the original signals in fixed magnitudes. However, the source signals change their local average magnitudes in many practical applications such as evoked potentials and speech by humans. In such cases, when a component source signal becomes very small suddenly, the corresponding row of the separating matrix tends to be large in the learning process to compensate for this change and to emit the output signal of the same amplitude. In particular, when one source signal becomes silent, the separating matrix diverges. This causes serious numerical instability.

This article proposes an efficient learning algorithm with equivariant properties that works well even when the magnitudes of the source signals change quickly. To this end, we first introduce an equivalence relation in the set of all the demixing matrices. Two matrices are defined to be equivalent when the recovered signals differ only by scaling. When the trajectory of the learning dynamics of the demixing matrix moves in a single equivalence class, the recovered signals remain the same except for scaling. Therefore, such a motion is ineffective in learning.

In order to reduce ineffective movements in the learning process, we modify the natural gradient learning rule (Amari, Cichocki, & Yang, 1996; Amari, 1998) such that the trajectory of learning is orthogonal to the equivalence class. This imposes constraints on the directions of movements in learning. However, such constraints are proved to be nonholonomic, implying that there are no integral submanifolds corresponding to the constraints. This implies that the trajectory can reach any demixing matrices in spite of the constraints, a property that is useful in robotic design. Nonholonomic dynamics is a current research topic in robotics.

Since the trajectory governed by the proposed algorithm does not include useless (redundant) components, the algorithm can be more efficient (with respect to convergence and efficiency) than those algorithms with other
constraints. It is also robust against sudden changes in the magnitudes of the source signals. Even when some components disappear, the algorithm still works well. Therefore, our algorithm can be used even when the number of source signals is overestimated. These features have been confirmed by computer simulations. The stability conditions for the proposed algorithm are also derived.

2 Critical Review of Existing Adaptive Learning Algorithms

The problem of blind source separation is formulated as follows (Hérault & Jutten, 1986; Jutten & Hérault, 1991). Let $n$ source signals $s_1(t), \ldots, s_n(t)$ at time $t$ be represented by a vector $s(t) = [s_1(t) \cdots s_n(t)]^T$, where the superscript $T$ denotes transposition of a vector or matrix. They are assumed to be stochastically independent. We assume that the source signals are not directly accessible, and we can instead observe $n$ signals $x_1(t), \ldots, x_n(t)$ summarized in a vector $x(t) = [x_1(t) \cdots x_n(t)]^T$, which are instantaneous linear mixtures of the $n$ source signals. Therefore, we have

$$x(t) = As(t), \quad (2.1)$$

where $A = [a_{ij}]_{n \times n}$ is an unknown nonsingular mixing matrix. The source signals $s(1), s(2), \ldots, s(T)$ are generated at discrete times $t = 1, 2, \ldots, T$ subject to unknown probability distributions. In order to recover the original source signals $s(t)$, we make new linear mixtures of the sensor signals $x(t)$ by

$$y(t) = W(t)x(t), \quad (2.2)$$

where $W = [w_{ij}]_{n \times n}$ is called the demixing or separating matrix. If $W(t) = A^{-1}$, the original source signals are recovered exactly. However, $A$ or $W$ is unknown, so we need to estimate $W$ from the available observations $x(1), \ldots, x(t)$. Here, $W(t)$ is an estimate.

It is well known that the mixing matrix $A$, and hence source signals, are not completely identifiable. There are two types of indeterminacies or ambiguities: the order of the original source signals is indefinite, and the scales of the original signals are indefinite. The first indeterminacy implies that a permutation or reordering of recovered signals remains possible, but this is not critical, because what is wanted is to extract $n$ independent signals; their ordering is not important. The second indeterminacy can be explained as follows. When the signals $s$ are transformed as $s' = As$, where $A$ is a diagonal scaling matrix with nonzero diagonal entries, the observed signals are kept invariant if the mixing matrix $A$ is changed into $A' = AA^{-1}$, because $x = As = A's'$. This implies that $W$ and $W' = AW$ are equivalent in the sense that they give the same estimated source signals except for their scales. Taking these indeterminacies into account, we can estimate
$W = A^{-1} P \Lambda$ at best, where $P$ is a permutation matrix and $\Lambda$ is a diagonal matrix.

These intrinsic (unavoidable) ambiguities do not seem to be a severe limitation, because in most applications, most information is contained in the waveforms of the signals (for instance, speech signals and multisensor biomedical recordings) rather than their magnitudes and the order in which they are arranged.

In order to make the scaling indeterminacies clearer, we introduce an equivalence relation in the $n^2$-dimensional space of nonsingular matrices $GL(n) = \{W\}$. We define $W$ and $\Lambda W$ as equivalent,

$$W \sim \Lambda W,$$

where $\Lambda$ is an arbitrary nonsingular diagonal matrix. Given $W$, its equivalence class

$$C_W = \{W' | W' = \Lambda W\}$$

is an $n$-dimensional subspace of $GL(n)$ consisting of all the matrices equivalent to $W$. Therefore, a learning algorithm need not search for the nonidentifiable $W = A^{-1}$ but searches for the equivalence class $C_W$ that contains the true $W = A^{-1}$ except for permutations.

In existing algorithms, in order to determine the scales, one of two constraints is imposed explicitly or implicitly.

The first is a rigid constraint. Let us introduce $n$ restrictions on $W$,

$$f_i(W) = 0, \quad i = 1, \ldots, n.$$  \hspace{1cm} (2.5)

A typical one is

$$f_i(W) = w_{ii} - 1, \quad i = 1, \ldots, n.$$  \hspace{1cm} (2.6)

Define the set $Q_r$ of $W$'s satisfying the above constraints to be

$$Q_r = \{W | f_i(W) = 0, \quad i = 1, \ldots, n\}.$$  \hspace{1cm} (2.7)

The set $Q_r$ is supposed to form a $(n^2 - n)$-dimensional manifold transversal to $C_W$. In order to resolve the scaling indeterminacies, we require that $W(t) = [w_{ij}(t)]$ always be included in the $Q_r$. Such a rigid constraint is implemented, for example, by the learning algorithm

$$w_{ij}(t + 1) = w_{ij}(t) - \eta(t)\psi_i(y_i(t))y_j(t), \quad i \neq j,$$

where $\psi(y_i)$ are suitably chosen nonlinear functions, $\eta(t)$ is the step size, and all the diagonal elements $w_{ii}$ are fixed equal to 1.
This idea of reducing the complexity of a learning algorithm slightly is a simple one and has been used in several papers (Héral & Jutten, 1986; Jutten & Héral, 1991; Sorouchyari, 1991; Chen & Chen, 1994; Moreau & Macchi, 1995; Matsuoka, Ohya, & Kawamoto, 1995; Oja & Karhunen, 1995). However, the constraints (see equation 2.5) do not allow equivariant properties (Cardoso & Laheld, 1996). Therefore, when $A$ is ill conditioned, the algorithm does not work well.

The second constraint specifies the power or energy of the recovered (estimated) signals. Such soft constraints can be expressed in a general form as

$$E[h_i(y_i)] = 0, \quad i = 1, \ldots, n,$$

(2.8)

where $E$ denotes the expectation and $h_i(y_i)$ is a nonlinear function; typically

$$h_i(y_i) = 1 - \varphi(y_i)y_i$$

(2.9)

or

$$h_i(y_i) = 1 - y_i^2.$$  

(2.10)

The latter guarantees that the variances or powers of the recovered signals $y_i$ are 1. Such soft constraints are implemented automatically in the adaptive learning rule,

$$W(t+1) = W(t) + \eta(t)F[y(t)]W(t),$$

(2.11)

where $F[y(t)]$ is an $n \times n$ matrix of the form (Cichocki, Unbehauen, & Rummert, 1994; Amari et al., 1996)

$$F[y(t)] = I - \varphi(y(t))y^T(t),$$

(2.12)

(or Amari et al., 1997)

$$F[y(t)] = I - \alpha \varphi(y(t))y^T(t) + \beta y(t)\varphi^T(y(t))$$

(2.13)

with adaptively or adequately chosen constants $\alpha$ and $\beta$, or (Cardoso & Laheld, 1996),

$$F[y(t)] = I - y(t)y^T(t) - \varphi(y(t))y^T(t) + y(t)\varphi(y^T(t)).$$

(2.14)

The diagonal elements of the matrix $F(y)$ in equation 2.11 can be regarded as specifying the constraints,

$$E[h_i(y_i)] = 0.$$  

(2.15)
Then, although $W(t)$ can move freely in $C_W$ in the learning process, any equilibrium point of the averaged learning equation satisfies the constraints of equation 2.15.

Let us define the set $Q_s$ of $W$'s by

$$Q_s = \{ W | E[h_i(y_i)] = 0, \quad i = 1, \ldots, n \}.$$  \hspace{1cm} (2.16)

Then $Q_s$ is an $(n^2 - n)$-dimensional attracting submanifold and $W(t)$ is attracted toward $Q_s$, although it does not stay exactly in $Q_s(t)$ because of stochastic fluctuations.

Summarizing, rigid constraints do not permit algorithms with equivariant properties (Cardoso & Laheld 1996). Soft constraints enable us to construct algorithms with equivariant properties. However, for source signals whose amplitudes or powers change over time (e.g., signals with long silent periods and spikes or rapid jumps like speech signals), it is more difficult to achieve rapid convergence to equilibria because of inadequate specification of constraints than in the stationary case. Special tricks or techniques may be employed, like random selection of training samples and dilation of silence periods (Bell & Sejnowski, 1995; Amari, Douglas, & Cichocki, 1998). Performance may be poorer for such nonstationary signals. Moreover, it is easy to show that such algorithms 2.11 through 2.14 are unstable (in the sense that weights $w_{ij}(t)$ diverge to infinity) in cases when the number of the true sources is fewer than the number of the observed signals, where $W$ is a quadratic $n \times n$ matrix. Such a case arises often in practice when one or more source signals are silent or decay to zero in an unpredictable way.

In this article, we will elucidate the roles of such constraints geometrically. We then show that such constraints and drawbacks can be lifted or at least relaxed by introducing a new algorithm. In other words, we impose no explicit constraints for the separating matrix $W$ and search for the equivalence class $C_W$ directly. This article is theoretical in nature, and simulations are limited. Larger-scale simulations studies and applications to real data are necessary for proving the practical usefulness of the proposed method.

3 Derivation of a New Learning Algorithm

In order to derive a new algorithm, we recapitulate the derivation of the natural gradient algorithm due to Amari, Chen, & Cichocki (1997). To begin, we consider the loss function,

$$l(y, W) = -\log |\det(W)| - \sum_{i=1}^{m} \log p_i(y_i),$$  \hspace{1cm} (3.1)

where $p_i(y_i)$ are some positive probability density functions (p.d.f.) of output.
signals and \( \det(W) \) denotes the determinant of matrix \( W \). We can then apply the stochastic gradient descent learning method to derive a learning rule.

In order to calculate the gradient of \( l \), we derive the total differential \( dl \) of \( l \) when \( W \) is changed from \( W \) to \( W + dW \). In component form,

\[
dl = l(y, W + dW) - l(y, W) = \sum_{i,j} \frac{\partial l}{\partial w_{ij}} dw_{ij},
\]

(3.2)

where the coefficients \( \frac{\partial l}{\partial w_{ij}} \) of \( dw_{ij} \) represent the gradient of \( l \). Simple algebraic and differential calculus yields

\[
dl = -\text{tr}(dWW^{-1}) + \varphi(y)^T dy,
\]

(3.3)

where \( \text{tr} \) is the trace of a matrix and \( \varphi(y) \) is a column vector whose components are

\[
\varphi(y_i) = -\frac{d \log(p_i(y_i))}{dy_i} = -\frac{\dot{p}_i(y_i)}{p_i(y_i)},
\]

(3.4)

\( \cdot \) denoting differentiation. From \( y = Wx \), we have

\[
dy = dWx = dWW^{-1} y.
\]

(3.5)

Hence, we define

\[
dx = dWW^{-1},
\]

(3.6)

whose components \( dx_{ij} \) are linear combinations of \( dw_{ij} \). The differentials \( dx_{ij} \) form a basis of the tangent space at \( W \) of \( \text{Gl}(n) \), since they are linear combinations of the basis \( dw_{ij} \).

The space \( \text{Gl}(n) \) of all the nonsingular matrices is a Lie group. An element \( W \) is mapped to the identity \( I \) when \( W^{-1} \) is multiplied from the right. A small deviation \( dW \) at \( W \) is mapped to \( dX = dWW^{-1} \) at \( I \). By using \( dX \), we can compare deviations \( dW_1 \) and \( dW_2 \) at two different points \( W_1 \) and \( W_2 \), using \( dX_1 = dW_1W_1^{-1} \) and \( dX_2 = dW_2W_2^{-1} \). The magnitude of deviation \( dX \) is measured by the inner product

\[
\langle dX, dX \rangle_I = \text{tr} \left( dX dX^T \right) = \sum_{i,j} (dx_{ij})^2.
\]

(3.7)

This implies that the inner product of \( dW \) at \( W \) is given by

\[
\langle dW, dW \rangle_W = \text{tr} \left\{ dWW^{-1} \left( dWW^{-1} \right)^T \right\}.
\]

(3.8)
which is different from $\sum (dw_i)^2$. Therefore, the space $Gl(n)$ has a Riemannian metric. It should be noted that $dX = dWW^{-1}$ is a nonintegrable differential form so that we do not have a matrix function $X(W)$ which gives equation 3.6. Nevertheless, the nonholonomic basis $dX/\gamma$ is useful. Since the basis $dX$ is orthonormal in the sense that $\|dX\|^2$ is given by equation 3.7, it is effective to analyze the learning equation in terms of $dX$. The natural Riemannian gradient (Amari, Cichocki, & Yang, 1995; Amari et al., 1996, 1998) is automatically implemented by it and the equivariant properties automatically hold. It is easy to rewrite the results in terms of $dW$ by using equation 3.6.

The gradient $dl$ in equation 3.3 is expressed in the differential form:

$$dl = -tr(dX) + \varphi(y)^T dX y.$$  (3.9)

This leads to the stochastic gradient learning algorithm,

$$\Delta X(t) = X(t + 1) - X(t) = -\eta(t) \frac{dl}{dX} = \eta(t) \left[ I - \varphi(y(t))y^T(t) \right].$$  (3.10)

in terms of $\Delta X(t) = \Delta W(t)W^{-1}(t)$. This is rewritten as

$$W(t + 1) = W(t) + \eta(t) \left[ I - \varphi(y(t))y^T(t) \right] W(t)$$  (3.11)

in terms of $W(t)$.

It was shown that the diagonal term in equation 3.11 can be set arbitrarily (Cichocki, Unbehauen, Moszczynski, & Rummert, 1994; Cichocki & Unbehauen, 1996; Amari & Cardoso, 1997). The problem has been analyzed also based on the information geometry of semiparametric statistical models (Amari & Kawanabe, 1997a, 1997b; Bickel, Klaassen, Ritov, & Wellner, 1993). Therefore, the above algorithm can be generalized in a more flexible and universal form as

$$W(t + 1) = W(t) + \eta(t) \left[ \Lambda(t) - \varphi(y(t))y^T(t) \right] W(t),$$  (3.12)

where $\Lambda(t)$ is any positive definite scaling diagonal matrix. By determining $\Lambda$ correctly depending on $y$, we have various algorithms. We propose

$$\Lambda(t) = \text{diag} \left\{ y_1(t)\varphi_1(y_1(t)), \ldots, y_n(t)\varphi_n(y_n(t)) \right\}.$$  

This algorithm is effective when the nonlinear activation functions $\varphi_i(y_i)$ are suitably chosen.

This corresponds to setting new constraints

$$dx_i(t) = 0, \quad i = 1, \ldots, n.$$  (3.13)
Nonholonomic Orthogonal Learning Algorithms

The new constraints lead to a new learning algorithm, written as

\[ W(t + 1) = W(t) - \eta(t)F[y(t)]W(t), \]  

where all the elements on the main diagonal of the \( n \times n \) matrix \( F(y) \) are put equal to zero, that is,

\[ f_{ii} = 0 \]  

and

\[ f_{ij} = \psi_i(y_i)y_j, \quad i \neq j. \]  

The learning rule, 3.14, is called the orthogonal nonholonomic natural gradient descent algorithm for the reason shown in the next section.

Algorithm 3.14 can be generalized as follows (Amari et al. 1997),

\[ W(t + 1) = W(t) - \eta(t)F[y(t)]W(t), \]  

with entries of the matrix \( F(y) \) as

\[ f_{ij} = \delta_{ij}\lambda_{ii} + \alpha_1 y_i y_j + \alpha_2 \psi_i(y_i)y_j - \alpha_3 y_i \psi_i(y_i). \]  

where \( \delta_{ij} \) is Kronecker delta, \( \lambda_{ii} = -\alpha_1 y_i^2 - \alpha_2 \psi_i(y_i)y_i + \alpha_3 y_i \psi_i(y_i) \) and \( \alpha_1, \alpha_2, \) and \( \alpha_3 \) are parameters that may be determined adaptively.

4 Nonholonomicity and Orthogonality

The space \( \text{Gl}(n) = \{W'\} \) is partitioned into equivalence classes \( C_W \), where any \( W' \) belonging to the same \( C_W \) is regarded as an equivalent demixing matrix. Let \( dW_C \) be a direction tangent to \( C_W \), that is, \( W + dW_C \) and \( W \) belong to the same equivalence class.

The learning equation determines \( \Delta W(t) \) depending on the current \( y \) and \( W \). When \( \Delta W(t) \) includes components belonging to the tangent directions to \( C_W \), such components are ineffective because they drive \( W \) within the equivalent class. Therefore, it is better to design a learning rule such that its trajectories \( \Delta W(t) \) are always orthogonal to the equivalence classes. Since \( C_W \) are \( n \)-dimensional subspaces, if we could find a family of \( n^2 - n \)-dimensional submanifolds \( Q \)'s such that \( C_W \) and \( Q \) are orthogonal, we could impose the constraints that the learning trajectories would belong to \( Q \). Therefore, one interesting question arises: Is there a family of \( (n^2 - n) \)-dimensional submanifolds \( Q \) that are orthogonal to the equivalence classes \( C_W \)? The answer is no. We can prove now that there does
not exist a submanifold that is orthogonal to the families of submanifolds $C_W$'s.

**Theorem 1.** The direction $dW$ is orthogonal to $C_W$, if and only if

$$dx_{ii} = 0, \quad i = 1, \ldots, n$$

where $dX = dWW^{-1}$.

**Proof.** Since the equivalent class $C_W$ consists of matrices $\Lambda W$, $\Lambda$ is regarded as a coordinate system in $C_W$. A small deviation of $W$ in $C_W$ is written as

$$dW_C = d\Lambda W,$$  

where $d\Lambda$ is diag($d\lambda_1, \ldots, d\lambda_n$). The tangent space of $C_W$ is spanned by them. The inner product of $dW$ and $dW_C$ is given by

$$\langle dW, dW_C \rangle_W = \langle dWW^{-1}, dW_C W^{-1} \rangle_I$$

$$= \langle dX, d\Lambda \rangle_I = \sum_{i,j} dx_{ij} d\lambda_{ij}$$

$$= \sum dx_{ij} d\lambda_{ij}.$$

Therefore, $dW$ is orthogonal to $C_W$ when and only when $dW$ satisfies $dx_{ii} = 0$ for all $i$.

We next show that $dX$ is not integrable, that is, there are no matrix functions $G(W)$ such that

$$dX = \frac{\partial G}{\partial W} \circ dW,$$  

where

$$\frac{\partial G}{\partial W} \circ dW = \sum_{i,j} \frac{\partial G}{\partial w_{ij}} dw_{ij}. $$

If such $G$ exists, $X = G(W)$ defines another coordinate system in $Gl(n)$. Even when such $G$ does not exist, $dX$ is well defined, and it forms a basis in the tangent space of $Gl(n)$ at $W$. Such a basis is called a nonholonomic basis (Schouten, 1954; Frankel, 1997).

**Theorem 2.** The basis defined by $dX = dWW^{-1}$ is nonholonomic.
Nonholonomic Orthogonal Learning Algorithms

**Proof.** Let us assume that there exists a function $G(W)$ such that

$$dX = dG(W) = dWW^{-1}.$$  \hspace{1cm} (4.6)

We now consider another small deviation from $W$ to $W + \delta W$. We then have

$$\delta dX = \delta dG = dW \delta(W^{-1}).$$  \hspace{1cm} (4.7)

We have

$$\delta(W^{-1}) = (W + \delta W)^{-1} - W^{-1} \approx -W^{-1}\delta WW^{-1}.$$  \hspace{1cm} (4.8)

Hence,

$$\delta dX = -dWW^{-1}\delta WW^{-1}.$$  \hspace{1cm} (4.9)

Since matrices are in general noncommutative, we have

$$\delta dX \neq d\delta X.$$  \hspace{1cm} (4.10)

This shows that there does not exist a matrix $X$ because $d\delta G = \delta dG$ always hold when a matrix $G$ exists.

Our constraints

$$dx_{ii} = 0, \quad i = 1, \ldots, n$$  \hspace{1cm} (4.11)

restrict the possible directions of $\Delta W$ and define $(n^2 - n)$-dimensional movable directions at each point $W$ of $Gl(n)$. These directions are orthogonal to $C_W$. However, by the same reasoning, there do not exist functions $g_i(W)$ such that

$$dx_{ii} = d\tilde{g}_i(W) = \sum_{j,k} \frac{\partial g_i(W)}{\partial w_{jk}} dw_{jk}.$$  \hspace{1cm} (4.12)

This implies that there exist no subspaces defined by $g_i(W) = 0$.

Such constraints are said to be nonholonomic. The learning equation, 3.14, with equation 3.15 defines a learning dynamics with nonholonomic constraints. At each point $W$, $\Delta W$ is constrained in $(n^2 - n)$ directions. However, the trajectories are not constrained in any $(n^2 - n)$-dimensional subspace, and they can reach any points in $Gl(n)$.
This property is important when the amplitudes of $s_i(t)$ change over time. If the constraints are
\[ E[h_i(y_i)] = 0, \quad (4.13) \]
for example,
\[ E[y_i^2 - 1] = 0. \quad (4.14) \]
when $E[s_i^2]$ suddenly become 10 times smaller, the learning dynamics with the soft constraints make the $i$th row of $W$ 10 times larger in order to compensate for this change and keep $E[y_i^2] = 1$. Therefore, even when $W$ converges to the true $C_W$, large fluctuations emerge from the ineffective movements caused by changes in amplitude of $s_i$. This sometimes causes numerical instability. On the other hand, our nonholonomic dynamics is always orthogonal to $C_W$ so that such ineffective fluctuations are suppressed. This is confirmed by computer simulations a few of which are presented in section 6.

**Remark 1.** There is a great deal of research on nonholonomic dynamics. Classical analytical dynamics uses nonholonomic bases to analyze the spinning gyro (Whittaker, 1940). It was also used in relativity (Whittaker, 1940). Kron (1948) used the nonholonomic constraints to present a general theory of electromechanical dynamics of generators and motors. Nonholonomic properties play a fundamental role in continuum mechanics of distributed dislocations (see, e.g., Amari, 1962). An excellent explanation is found in Brockett (1993), where the controllability of nonlinear dynamical systems is shown by using the related Lie algebra. Recently robotics researchers have been eager to analyze dynamics with nonholonomic constraints (Suzuki & Nakamura 1995).

**Remark 2.** Theorem 2 shows that the orthogonal natural gradient descent algorithm evolves along a trajectory path that does not include redundant (useless) components in the directions of $C_W$. Therefore, it seems likely that the orthogonal algorithm can be more efficient than other algorithms, as has been confirmed by preliminary computer simulations.

5 Stability Analysis

In this section, we discuss the local stability of the orthogonal natural gradient descent algorithm, 3.14, defined by
\[ \Delta W(t) = -\eta(t)F[y(t)]W(t), \quad (5.1) \]
Theorem 3. When the zero-mean independent source signals satisfy conditions

\[ E(\phi_i(s_i)s_j) = 0, \quad \text{if} \quad i \neq j \]  

\[ E(\phi_i(s_i)s_k) = 0, \quad \text{if} \quad k \neq j \]  

and when the following inequalities hold,

\[ E(\phi_i(s_i)s_i^2)E(\phi_i(s_i)s_j^2) > E(s_i\phi_i(s_i))E(s_j\phi_i(s_j)), \quad i, j = 1, \ldots, n. \]  

then the desired equivalence class \( C_W, W = A^{-1}P\Lambda \) is a stable equilibrium. However, if conditions 5.2 and 5.3 are still satisfied and

\[ E(\phi_i(s_i)s_i^2)E(\phi_i(s_i)s_j^2) < E(s_i\phi_i(s_i))E(s_j\phi_i(s_j)), \]  

then replacing \( f_{ij} = \psi(y_i)y_j \) by

\[ f_{ij} = y_i\phi_i(y_j), \quad i \neq j \]  

it is guaranteed that the class \( C_W, W = A^{-1}P\Lambda \), contains stable separating equilibria.

Proof. Let \( W = A^{-1} \) and \((dW)_i\) denote the \( i \)th row of \( dW \) and \((W^{-1})^j\) be the \( j \)th column of \( W^{-1} \). Then taking into account that \( dx_{ij} = \psi_i(y_i)y_j \) and \( dx_{ii} = 0 \), we evaluate the second differentials in the Hessian matrix as follows:

\[ d^2x_{ij} = E[d(\phi_i(y_i)y_j)] \]  

\[ = E[\phi_i(y_i)y_idy_j] + E[\phi_i(y_i)dy_j] \]  

\[ = E[\phi_i(y_i)y_id(W^{-1}s)] + E[\phi_i(y_i)(dW)_j(W^{-1}s)] \]  

\[ = E[\phi_i(y_i)y_id(W^{-1})_i] + E[\phi_i(y_i)y_i(dW)_j(W^{-1})_i] \]  

\[ = -E[\phi_i(s_i)s_i^2]dx_{ij} - E[\phi_i(s_i)s_i]dx_{ji}. \]
Hence

\[ d^2x_{ij} = -E[\phi_i(s_i)s_i^2]dx_{ij} - E[\phi_i(s_i)s_i]dx_{ij} \]  

(5.8)

\[ d^2x_{ji} = -E[\phi_j(s_j)s_j^2]dx_{ji} - E[\phi_j(s_j)s_j]dx_{ji} \]  

(5.9)

\[ d^2x_{ii} = 0. \]  

(5.10)

By the assumptions made in theorem 3, we see that the equilibrium point \( W = A^{-1} \) is stable.

The proof for \( W = A^{-1}PA \) is the same and is omitted here for brevity. The proof for the other cases is similar.

From the proof of theorem 2, we have

**Corollary.** Let us consider the special case that all neurons have the same activation function: the cubic nonlinearity \( \phi(y_i) = y_i^3 \). In this case, if all the source signals are subgaussian signals (with negative kurtosis), then the learning algorithm, 3.14, with \( f_{ii} = 0 \) and \( f_{ij} = y_i^3y_j \), if \( i \neq j \) ensures local stability for the desired equilibrium points: \( W = A^{-1}PA \). Analogically, if all the source signals are supergaussian (with positive kurtosis), then the algorithm, 3.14, with \( f_{ii} = 0 \) and \( f_{ij} = y_iy_j^3 \), if \( i \neq j \) ensures that all the desired equilibrium points \( W = A^{-1}PA \) are stable equilibrium separating points.

6 Computer Simulations

In our computer experiments we have assumed that source signals have generalized gaussian distributions, and corresponding nonlinear activation functions have the form \( \phi(y_i) = |y_i|^{r_i - 1} \text{sgn}(y_i) \) for \( r_i \geq 1 \) and \( \phi(y_i) = y_i/[(y_i^{2-r_i}) + \epsilon] \) for \( 0 \leq r_i < 1 \), where \( \epsilon \) is small positive constant added to prevent singularity of the nonlinear function (Cichocki, Bogner, Moszczynski, & Pope, 1997; Cichocki, Sabala, & Amari, 1998). For spiky supergaussian sources the optimal parameter \( r_i \) changes between zero and one. In the illustrative examples that follow, we have assumed that \( r_i = 0.5 \) and the learning rate is fixed equal to \( \eta = 0.001 \).

Preliminary computer simulation experiments have confirmed the high performance of the proposed algorithm. We will give two illustrative examples to show the efficiency of the orthogonal natural gradient adaptive algorithm and verify the validity of our theorems. We also show that the orthogonal natural gradient descent algorithm is usually more efficient than soft constraint natural gradient descent algorithms for nonstationary
signals, especially when the number of sources changes dynamically in time.

6.1 Example 1. To show the effectiveness of our algorithm, we assume that the number of sources is unknown (but their number is not greater than the number of the sensors). Three natural speech signals shown in Figure 1a are mixed in four sensors by using an ill-conditioned $4 \times 3$ mixing matrix. It is assumed that the mixing matrix and the number of the sources are unknown. It should be noted that the sensor signals look almost identical (see Figure 1b). Recovered signals are shown in Figure 2a.
The algorithm was able after 5000 iterations not only to estimate the waveform of original speech signals but also their number. It should be noted that one output signal quickly decays to zero. For comparison we applied the standard independent component analysis learning algorithm, 3.11, to the same noiseless data. The standard algorithm also recovered the original three signals well, but it creates a false signal, as shown in Figure 2b. Moreover, the norm of separating matrix $W$ is increasing exponentially, as illustrated in Figure 3a. This problem can be circumvented by computing preliminarily the rank of covariance matrix of the sensor signals and then taking only the number of sensors equal to the rank, provided noise
Figure 3: Convergence dynamic behavior of the Frobenious norm of a separating matrix $W$ as a function of the number of iterations for (a) the standard algorithm (3.11) for noiseless case, (b) the standard algorithm (3.11) for 1% additive gaussian noise, and (c) the orthogonal natural gradient descent algorithm (3.14).
is negligible. However, such an approach is not very practical for a nonstationary environment when the number of sources changes continuously in time.

When some small (1%) additive noise is added to sensors, the standard algorithm can be stabilized, as shown in Figure 3b. However, a fluctuation of the norm of separating matrix $W$ is larger than for the orthogonal natural gradient descent algorithm (see Figures 3b and 3c).

6.2 Example 2. In this simulation, we separated seven highly nonstationary natural speech signals, as shown in Figure 4a, using seven sensors (microphones), assuming that two of sources are temporally zero (silence period). It is assumed again that the number of active source signals and the randomly chosen mixing matrix $A$ are completely unknown. The new learning algorithm, 3.14, was able to separate all signals, as illustrated in Figure 4c. Again we have found by simulation that the standard learning algorithm, 3.11, demonstrates higher fluctuation of synaptic weights of the separating matrix. Moreover, the absolute values of the weights increase exponentially to infinity for the noiseless case, so we are not able to compare the performance of the two algorithms, because the standard ICA algorithm, 3.11, is principally unstable in the noiseless case (without computing the rank of covariance matrix of sensor signals). Even for small noise, the weights of the separating matrix take very high values and are unstable in the case of the standard ICA method.

7 Conclusion

We have pointed out several drawbacks in existing adaptive algorithms for blind separation of signals—for example, the problem of separation of nonstationary signals and the problem of stability if some sources decay to zero. To avoid these problems, we developed a novel orthogonal natural gradient descent algorithm with the equivariant property. We have demonstrated by computer simulation experiments that the proposed algorithm ensures a good convergence speed even for ill-conditioned problems. The algorithm solves the problem of on-line estimation of the waveforms of the source signals even when the number of sensors is larger than the number of sources that is unknown. Moreover, the algorithm works properly and precisely for highly nonstationary signals like speech signals and biomedical signals. It is proved theoretically that the trajectories governed by the algorithm do not include redundant (useless) components and the algorithm is potentially more efficient than algorithms with soft or rigid constraints. The main feature of the algorithm is its orthogonality and that it converges to an equivalent class of the separating matrices rather than isolated critical points. Local stability conditions for the developed algorithm are also given.
Figure 4: Computer simulation for example 2. (a) Original active speech signals. (b) Mixed (microphone) signals. (c) Separated (estimated) source signals.
References


---

Received September 15, 1997; accepted June 4, 1998.