The Nonlinear PCA Criterion in Blind Source Separation: Relations with Other Approaches

Juha Karhunen, Petteri Pajunen, and Erkki Oja

Helsinki University of Technology, Laboratory of Computer and Information Science

P.O.Box 2200, FIN-02015 HUT, Espoo, FINLAND

email: Juha.Karhunen@hut.fi, Petteri.Pajunen@hut.fi, Erkki.Oja@hut.fi

Fax: +358-9-451 3277, http://www.cis.hut.fi

Abstract

We present new results on the nonlinear PCA (Principal Component Analysis) criterion in blind source separation (BSS). We derive the criterion in a form that allows easy comparisons with other BSS and Independent Component Analysis (ICA) contrast functions like cumulants, Bussgang criteria, and information theoretic contrasts. This clarifies how the nonlinearity should be chosen optimally. We also discuss the connections of the nonlinear PCA learning rule with the Bell-Sejnowski algorithm and the adaptive EASI algorithm. Furthermore, we show that a nonlinear PCA criterion can be minimized using least-squares approaches, leading to computationally efficient and fast converging algorithms. The paper shows that nonlinear PCA is a versatile starting point for deriving different kinds of algorithms for blind signal processing problems.

Keywords: blind separation, nonlinear PCA, least squares, unsupervised learning, neural networks

1 Introduction

In blind source separation, the goal is to recover mutually independent but otherwise unknown source signals from their linear mixtures without knowing the mixing coefficients. Such blind techniques have many applications especially in signal and image processing; for a recent review, see [15].

We consider the standard signal model used in BSS, which is [11, 14, 2]

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t). \tag{1}$$

Here $\mathbf{s}(t) = [s_1(t), \ldots, s_m(t)]^T$ is the source vector at time t. Its components are the m unknown but mutually independent source signals $s_i(t)$, $i = 1, \ldots, m$. The components of the m-dimensional data vector $\mathbf{x}(t)$ are some linear mixtures of the source signals. The $m \times m$ mixing matrix $\mathbf{A}$ is an unknown full rank constant matrix. At most one of the source signals $s_i(t)$ is allowed to be Gaussian.

Recently, there has been a lot of interest in BSS problems both in statistical signal processing and unsupervised neural learning. Usually some higher-order statistics must be utilized for achieving separation. In adaptive and neural learning rules [11, 14, 2, 25, 1, 5, 17, 8, 22], this is typically done implicitly by using some suitable nonlinearities during the learning phase.

We have studied nonlinear PCA type methods, and shown that these simple neural learning rules can be successfully applied to the BSS problem [17, 22]. Up to now, their connections to information-theoretic contrasts have not been known. In this paper we show that a nonlinear PCA criterion, when the nonlinear function is chosen suitably, is intimately related to both cumulant contrasts, Bussgang criteria, and maximum likelihood criteria. It can be used for deriving algorithms of different type and complexity which have close relationships to some other existing and widely used algorithms. The emphasis is here on theoretical derivations and on showing connections between various contrasts and algorithms.

2 The nonlinear PCA criterion

2.1 The basic criterion and learning rule

There exist many possible nonlinear extensions of neural PCA learning, often leading to somewhat different solutions; see for example [21, 23, 12, 13]. In this paper, we concentrate on a specific form of nonlinear PCA which has recently turned out to be especially useful in blind source separation. This is obtained by minimizing the nonlinear PCA criterion [23, 12]

$$J_1(\mathbf{W}) = \mathrm{E}\{\|\mathbf{x} - \mathbf{W}\mathbf{g}(\mathbf{W}^T\mathbf{x})\|^2\} \tag{2}$$

with respect to the $m \times m$ weight matrix $\mathbf{W}$. More generally, $\mathbf{W}$ could be a rectangular matrix. Here $\mathbf{g}(\mathbf{y})$ denotes the vector which is obtained by applying a nonlinear function $g(t)$ componentwise to the vector $\mathbf{y}$. In the nonlinear PCA criterion (2), $g(t)$ is usually some odd function such as $g(t) = \tanh(t)$ or $g(t) = t^3$. The criterion (2) was first proposed in a more general context by Xu in [23].

The criterion (2) can be approximately minimized using the stochastic gradient descent algorithm

$$\Delta\mathbf{W} = \mu[\mathbf{x} - \mathbf{W}\mathbf{g}(\mathbf{W}^T\mathbf{x})]\mathbf{g}(\mathbf{x}^T\mathbf{W}). \tag{3}$$

This nonlinear PCA subspace learning rule has been independently derived in [23] and [12]. In (3), μ is a positive learning parameter, and $\Delta\mathbf{W}$ denotes the update of matrix $\mathbf{W}$. We have omitted the time index t from all the quantities for simplicity. The algorithm (3) was first proposed by Oja et al. in [21] on heuristic grounds. For more details and related algorithms, see [21, 23, 12, 13].

2.2 Application to blind source separation

In applying the nonlinear PCA criterion (2) and algorithm (3) to the BSS problem, it is essential that the data vectors $\mathbf{x}(t)$ are first prewhitened: let us denote the whitened input vectors by $\mathbf{v} = \mathbf{V}\mathbf{x}$. It holds that $E\{\mathbf{v}\mathbf{v}^T\} = \mathbf{I}$. Thus the nonlinear PCA learning rule is applied to BSS problems in the form

$$\Delta\mathbf{W} = \mu[\mathbf{v} - \mathbf{W}\mathbf{g}(\mathbf{y})]\mathbf{g}(\mathbf{y}^T) \tag{4}$$

where

$$\mathbf{y} = \mathbf{W}^T\mathbf{v} = \mathbf{W}^T\mathbf{V}\mathbf{x} = \mathbf{B}\mathbf{x}. \tag{5}$$

Here $\mathbf{y}$ is the output vector. When the algorithm has converged to the solution, $\mathbf{y}$ should contain as its elements estimates of the sources $s_i(t)$, possibly in different order and scale. If we assume that the variances of the $s_i(t)$ are equal to 1, then we may require that after convergence $\mathrm{E}\{\mathbf{y}\mathbf{y}^T\} = \mathbf{I}$. Note that this implies that $\mathbf{W}$ is an orthogonal matrix: from (5), $\mathrm{E}\{\mathbf{y}\mathbf{y}^T\} = \mathbf{I} = \mathbf{W}^T\mathrm{E}\{\mathbf{v}\mathbf{v}^T\}\mathbf{W} = \mathbf{W}^T\mathbf{W}$. The total separating matrix between input and output vectors is $\mathbf{B} = \mathbf{W}^T\mathbf{V}$.

We have shown [22, 17] that for the prewhitened mixture vectors $\mathbf{v}(t)$, an asymptotically stable solution for $\mathbf{W}(t)$ in (4) is an $m \times m$ separating matrix with orthogonal columns, provided that all the source signals are of the same type; either sub-Gaussian or super-Gaussian. Although global convergence to a separating matrix has not been proven, numerical simulations show that the algorithm (4) usually converges to the desired solution. In order to achieve separation, it suffices that the nonlinearity $g(t)$ is of the right type [3, 17]. We have used in our earlier experiments the sigmoidal nonlinearity $g(t) = \tanh(t)$ for sub-Gaussian sources with good results. The nonlinear PCA rule (4) can be applied also for super-Gaussian sources using Fahlman type activation functions [8]. The separation properties of the algorithm (4) have been analyzed rigorously in the general case in [22].

For prewhitened data vectors $\mathbf{v}$, the orthogonality of the separating matrix allows us to analyze the criterion (2) in more detail. Let us consider the problem of minimizing the cost function

$$J_1(\mathbf{W}) = \mathrm{E}\{\|\mathbf{v} - \mathbf{W}\mathbf{g}(\mathbf{W}^T\mathbf{v})\|^2\}, \tag{6}$$

under the constraint

$$\mathbf{W}^T\mathbf{W} = \mathbf{I}. \tag{7}$$

Note that this constrained problem is not the same as the problem of minimizing (6) without any constraints on matrix $\mathbf{W}$. The nonlinear PCA learning rule (4) was derived for the unconstrained problem. However, it turns out that the solution $\mathbf{W}$ given by the nonlinear PCA learning rule has orthogonal columns, and $\mathbf{W}^T\mathbf{W} = \mathbf{D}$, a diagonal matrix. The diagonal elements depend on the nonlinearity g and the densities of the sources s_i. In the following, we assume for clarity that all the diagonal elements are the same and in fact equal to one, or $\mathbf{D} = \mathbf{I}$. If all the sources have the same density, this is achieved by a simple adjustment of the g function. If the assumption does not hold, then the diagonal elements of $\mathbf{D}$ enter into the derived results in an obvious way.

Assuming (7), we obtain

$$\begin{aligned} J_1(\mathbf{W}) &= \mathrm{E}\{\|\mathbf{v} - \mathbf{W}\mathbf{g}(\mathbf{W}^T\mathbf{v})\|^2\} \\ &= \mathrm{E}\{\|\mathbf{W}^T\mathbf{v} - \mathbf{W}^T\mathbf{W}\mathbf{g}(\mathbf{W}^T\mathbf{v})\|^2\} \\ &= \mathrm{E}\{\|\mathbf{y} - \mathbf{g}(\mathbf{y})\|^2\} \\ &= \sum_{i=1}^{m} \mathrm{E}\{[y_i - g(y_i)]^2\}. \end{aligned} \tag{8}$$

This is the key result relating the Nonlinear PCA criterion to other contrast functions proposed for the BSS problem. The simple form $\sum_{i=1}^{m} \mathrm{E}\{[y_i - g(y_i)]^2\}$ makes it easier to understand the effect of any nonlinearity to separation, and allows a comparison to known contrast functions [6].

3 Relationship of the Nonlinear PCA criterion with other approaches

Many contrasts like Bussgang, maximum likelihood, negentropy, etc. in the case of an orthogonal separating matrix lead to a criterion of the form

$$J(\mathbf{W}) = \sum_{i=1}^{m} \mathrm{E}\{f(y_i)\} \tag{9}$$

with $f(y_i)$ a nonlinear function, depending on the criterion. For example, in maximum likelihood estimation the function $f(y_i)$ would be $-\log p_i(y_i)$, where $p_i(y_i)$ is the probability density of y_i; in cumulant contrasts, $f(y_i)$ would be a polynomial. If function $g(y_i)$ in (8) is chosen such that

$$[y_i - g(y_i)]^2 = f(y_i), \tag{10}$$

then the Nonlinear PCA criterion becomes equivalent to (9). Note that the assumption made here that functions g and f are the same in each term of the sum is not necessary; we could replace $g(y_i)$ in (8) by $g_i(y_i)$ to allow for more general cases. However, for clarity, we consider here only the case that there is one nonlinear function $g(y_i)$.

Some concrete examples on the relations are given next.

3.1 Relationship with cumulant-based contrasts

As the first example, note that for the odd quadratic function

$$g(y) = \begin{cases} y^2 + y, & y \geq 0; \\ -y^2 + y, & y < 0. \end{cases} \tag{11}$$

the criterion (8) becomes

$$J_1(\mathbf{W}) = \sum_{i=1}^{m} \mathrm{E}\{[y_i - y_i \pm y_i^2]^2\} = \sum_{i=1}^{m} \mathrm{E}\{y_i^4\} \tag{12}$$

The statistic $J_1(\mathbf{W}) = \sum_i \mathrm{E}\{y_i^4\}$ has been rigorously shown to be a contrast in [20] under the following assumptions: all the sources have the same known sign of kurtosis, and the data have been prewhitened. Therefore, all the global minima (there are several) of J_1 correspond exactly to the separating solutions. In experiments, local optimization of J_1 usually provides the separated sources.

3.2 Relationship with Bussgang criterion and entropy-based contrasts

The form $\mathrm{E}\{\|\mathbf{y} - \mathbf{g}(\mathbf{y})\|^2\}$ is quite similar to the Bussgang cost function used in blind equalization [19, 9, 10]. We use Lambert's notation and approach [19], where the nonlinearity is chosen to be

$$g(y) = \frac{-\mathrm{E}\{|y|^2\} p'_y(y)}{p_y(y)}. \tag{13}$$

In (13), $p_y(y)$ is the density of y and $p'_y(y)$ its derivative.

In particular, Lambert [19] has given various algorithms for minimizing the cost function $\mathrm{E}\{\|\mathbf{y} - \mathbf{g}(\mathbf{y})\|^2\}$. However, Lambert's derivations are heuristic because he first derives algorithms for the LMS type cost function $\mathrm{E}\{\|\mathbf{y} - \mathbf{g}(\mathbf{y}^*)\|^2\}$, where $\mathbf{y}^*$ is a known reference vector. After this, the vector $\mathbf{y}^*$ is simply replaced by the unknown vector $\mathbf{y}$ for getting the respective algorithms for the blind Bussgang cost function. Thus exact gradients are in these algorithms approximated by essentially regarding the term

$\mathbf{g}(\mathbf{y})$ as a constant. In this manner, Lambert [19] ends up into the algorithm (using our notation)

$$\Delta \mathbf{W} = \mu \mathbf{v}[\mathbf{y} - \mathbf{g}(\mathbf{y})]^T \tag{14}$$

for minimizing the cost (8).

We can now give a rigorous derivation of Lambert's algorithm. In fact, the gradient algorithm (14) results from the information-theoretic cost function

$$J_2 = \sum_{i=1}^{m} E\{\log p_{y_i}(y_i) / \log p_G(y_i)\}, \tag{15}$$

which is the sum of the individual negentropies $J_{ni} = \mathrm{E}\{\log p_{y_i}(y_i) / \log p_G(y_i)\}$ of the components y_i of the output vector $\mathbf{y}$. In (15), $p_G(y_i)$ denotes the Gaussian density which has the same mean and variance as y_i, and p_{y_i} is the probability density of y_i. Negentropy measures the difference of the distribution p_{y_i} from the respective Gaussian distribution $p_G(y_i)$. In BSS applications we want to maximize (15).

The algorithm (14) can be derived from the criterion (15) as follows. Denoting by $\mathbf{w}_i$ the i-th column of the matrix $\mathbf{W}$ (i-th row of $\mathbf{W}^T$), the i-th output $y_i = \mathbf{w}_i^T \mathbf{v}$ depends only on the i-th weight vector $\mathbf{w}_i$. Thus

$$\frac{\partial J_2}{\partial \mathbf{w}_i} = \frac{\partial J_2}{\partial y_i} \frac{\partial y_i}{\partial \mathbf{w}_i} = \frac{\partial J_{ni}}{\partial y_i} \mathbf{v}. \tag{16}$$

Since $J_{ni} = \mathrm{E}\{\log p_{y_i}(y_i)$ - $\log c \exp(-y_i^2/(2\sigma_i^2))\}$ where σ_i^2 is the variance of y_i for zero-mean data, we get

$$\begin{aligned} \frac{\partial J_{ni}}{\partial y_i} &= \mathrm{E}\{p'_{y_i}(y_i)/p_{y_i}(y_i) + y_i/\sigma_i^2\} \\ &= 1/\sigma_i^2 \mathrm{E}\{y_i - g_i(y_i)\}, \end{aligned} \tag{17}$$

where $g_i(y_i) = -\sigma_i^2 p'_{y_i}(y_i)/p_{y_i}(y_i)$ is the Bussgang nonlinearity for y_i. Taking into account that $\sigma_i^2 = 1$ for prewhitened data yields

$$\frac{\partial J_2}{\partial \mathbf{w}_i} = \mathrm{E}\{(y_i - g_i(y_i))\mathbf{v}\}. \tag{18}$$

Thus the total gradient with respect to $\mathbf{W}$ becomes

$$\frac{\partial J_2}{\partial \mathbf{W}} = \mathrm{E}\{\mathbf{v}[\mathbf{y} - \mathbf{g}(\mathbf{y})]^T\}, \tag{19}$$

leading to the stochastic gradient algorithm (14).

Thus we can interpret the minimization of the cost $\mathrm{E}\{\|\mathbf{y} - \mathbf{g}(\mathbf{y})\|^2\}$ as minimizing the average squared norm of the instantaneous gradient vector $\partial J_2/\partial \mathbf{y} = \mathbf{y} - \mathbf{g}(\mathbf{y})$, and hence finding an extremal point of the sum of negentropies. These considerations relate the nonlinear PCA criterion for prewhitened data (8) closely to the negentropy criterion (15). The relationships of negentropy to other meaningful information theoretic principles (such as infomax or maximum likelihood) used for deriving constrast functions for the BSS problem have in turn been discussed in [4]. The results in [4] show that all these criteria or principles are closely connected on certain conditions.

4 Relationship of the nonlinear PCA learning rule with other algorithms

4.1 Relationship with Lambert's algorithm

In spite of similarities between the nonlinear PCA learning rule (4) and Lambert's learning rule (14), there are also some important differences which deserve attention. First, in our experiments the gradient algorithm (14) usually found only some of the source signals, while the remaining outputs were duplicates of the found source signals with a phase shift. An obvious explanation is that in the sum-of-negentropies criterion (15) there is nothing which forces the output signals $y_1, \ldots, y_m$ to be mutually different. Therefore, in the worst case all the outputs could converge to the same source signal without some extra restriction. On the other hand, the nonlinear PCA algorithm (4) forces the components of its output vector $\mathbf{y} = \mathbf{W}^T\mathbf{v}$ in practice mutually clearly different (even for nonwhite data), because otherwise the mean-square error $J_1(\mathbf{W})$ will not be minimized [12].

Another important point concerns the choice of appropriate nonlinearity $g(y_i)$. Because the outputs y_i should estimate the source signals s_j, ideally $g(y_i)$ should be matched to the density of the source signals $p(s_i)$. However, $p(s_i)$ is in practice usually unknown. Fortunately, it suffices that the nonlinearity $g(y_i)$ is of right type [17, 22, 3].

Inserting for example the often used sigmoidal nonlinearity $g(x) = \tanh(x)$ into the Bussgang formula (13) and assuming that $\mathrm{E}\{x^2\} = 1$ (for example after prewhitening) yields the solution $p_x(x) = 1/\cosh(x)$ for the arising differential equation. This means that (2) with $g(x) = \tanh(x)$ is a Bussgang blind equalization cost for sources with a density proportional to $p_x(x) = 1/\cosh(x)$. Now this density is super-Gaussian, and should therefore apply for source signals that have a positive kurtosis. While this in fact holds for the algorithm (14), for the nonlinear PCA algorithm (4) just the reverse is true: $\tanh(x)$ is a prototype nonlinearity for separating sub-Gaussian source signals.

This phenomenon can be explained by looking at the nonlinear Hebbian term $\mathbf{v}\mathbf{g}(\mathbf{y})^T$ which is mainly responsible for learning the separating solution in these algorithms. In (14), $\mathbf{v}\mathbf{g}(\mathbf{y})^T$ has a negative sign while its sign is positive in (4). Furthermore, if the outputs $\mathbf{y}$ are roughly independent after sufficient convergence has taken place then $\mathrm{E}\{\mathbf{g}(\mathbf{y})\mathbf{g}(\mathbf{y})^T\} \approx c\mathbf{I}$, where c is a constant and $\mathbf{I}$ the unit matrix.

4.2 Relationship with the Bell-Sejnowski algorithm

The nonlinear PCA rule (4) can be compared also to the well-known Bell-Sejnowski algorithm [1], which is in its basic form

$$\Delta\mathbf{B} = \mu[(\mathbf{B}^T)^{-1} - \mathbf{g}(\mathbf{y})\mathbf{x}^T] \tag{20}$$

If the data are prewhitened, $(\mathbf{B}^T)^{-1} = \mathbf{B}$. Using our notation, (20) then becomes

$$\Delta\mathbf{W} = \mu[\mathbf{W} - \mathbf{v}\mathbf{g}(\mathbf{y}^T)] \tag{21}$$

Comparing this with the nonlinear PCA rule (4) and assuming that there $\mathrm{E}\{\mathbf{g}(\mathbf{y})\mathbf{g}(\mathbf{y})^T\} \approx c\mathbf{I}$ shows that the update rule (21) becomes roughly the same but with reversed sign. The learning rule (20) has been derived by maximizing the joint entropy of nonlinearly transformed outputs $\mathbf{g}(\mathbf{y})$ with respect to the total separating matrix $\mathbf{B}$ [1, 25]. This criterion in turn has a close relationship to minimizing mutual information as discussed in [25]. The minimum of mutual information is achieved for independent outputs. Usually the algorithm (20) is applied in its natural gradient form, proposed by Amari, Cichocki, and Yang [25, 5]

$$\Delta \mathbf{B} = \mu[\mathbf{I} - \mathbf{g}(\mathbf{y})\mathbf{y}^T]\mathbf{B} \tag{22}$$

which avoids prewhitening and computation of the inverse matrix and converges much faster than the original algorithm (20).

Thus it is apparent that roughly speaking the nonlinear PCA rule is trying rather to maximize the sum of negentropies (15) than minimize it, or to minimize the joint entropy criterion instead of maximizing it. However, using an "opposite" type of nonlinearity this approach yields the desired result, independent components or estimates of the original source signals. In our earlier experiments [17] with the EASI algorithm [2] (which is discussed in the next subsection) it turned out that if the nonlinearity was indeed chosen so that it was of the opposite type, the algorithm tended to converge to a matrix $\mathbf{B}$ providing maximally mixed (mutually dependent) outputs.

A more direct relationship between the Amari - Cichocki - Yang algorithm (22) and nonlinear PCA learning rules can be obtained if an explicit whitening algorithm is introduced as follows. For the total separating matrix $\mathbf{B} = \mathbf{W}^T\mathbf{V}$, the update rule is generally

$$\Delta \mathbf{B} = \Delta \mathbf{W}^T \mathbf{V} + \mathbf{W}^T \Delta \mathbf{V}. \tag{23}$$

The second term $\mathbf{W}^T \Delta \mathbf{V}$ depends on the whitening algorithm. If the simple whitening algorithm

$$\Delta \mathbf{V} = [\mathbf{I} - \mathbf{v}\mathbf{v}^T]\mathbf{V} = [\mathbf{I} - \mathbf{V}\mathbf{x}\mathbf{x}^T\mathbf{V}^T]\mathbf{V} \tag{24}$$

is employed (the learning parameter is omitted here for convenience), one gets

$$\begin{aligned} \mathbf{W}^T \Delta \mathbf{V} &= \mathbf{B} - \mathbf{B}\mathbf{x}\mathbf{x}^T\mathbf{V}^T\mathbf{V} \\ &= \mathbf{B} - \mathbf{y}\mathbf{x}^T\mathbf{V}^T\mathbf{W}\mathbf{W}^T\mathbf{V} = [\mathbf{I} - \mathbf{y}\mathbf{y}^T]\mathbf{B}. \end{aligned} \tag{25}$$

Combining this with (23) and (22), and assuming that the learning rates for the whitening and separating algorithms are the same, we obtain (again omitting the learning parameter)

$$\Delta \mathbf{W}^T \mathbf{V} = \Delta \mathbf{B} - \mathbf{W}^T \Delta \mathbf{V} \tag{26}$$

$$= [\mathbf{I} - \mathbf{g}(\mathbf{y})\mathbf{y}^T]\mathbf{B} - [\mathbf{I} - \mathbf{y}\mathbf{y}^T]\mathbf{B} \tag{27}$$

$$= [\mathbf{y}\mathbf{y}^T - \mathbf{g}(\mathbf{y})\mathbf{y}^T]\mathbf{B}. \tag{28}$$

Multiplying with $\mathbf{V}^{-1}$ from the right yields

$$\Delta \mathbf{W}^T = [\mathbf{y}\mathbf{y}^T - \mathbf{g}(\mathbf{y})\mathbf{y}^T]\mathbf{W}^T, \tag{29}$$

or

$$\begin{aligned} \Delta \mathbf{W} &= \mathbf{W}[\mathbf{y}\mathbf{y}^T - \mathbf{y}\mathbf{g}(\mathbf{y}^T)] \\ &= -\mathbf{v}\mathbf{g}(\mathbf{y}^T) + \mathbf{W}\mathbf{y}\mathbf{y}^T. \end{aligned} \tag{30}$$

This is a simplified learning rule; comparing with (4), the sign has been changed and the term $\mathbf{W}\mathbf{g}(\mathbf{y})\mathbf{g}(\mathbf{y}^T)$ is replaced by $\mathbf{W}\mathbf{y}\mathbf{y}^T$. The nonlinearities are removed from this term. The learning rule can be further simplified by noting that $\mathbf{W}\mathbf{y}\mathbf{y}^T = \mathbf{W}\mathbf{W}^T\mathbf{v}\mathbf{y}^T = \mathbf{v}\mathbf{y}^T$; hence,

$$\Delta \mathbf{W} = -\mathbf{v}\mathbf{g}(\mathbf{y}^T) + \mathbf{v}\mathbf{y}^T \tag{31}$$

which is exactly Lambert's algorithm. The conclusion is that equivalently to the Amari - Cichocki - Yang learning rule (22), one can use a two-layer neural network in which the first layer uses the symmetrical whitening algorithm (24) and the second layer uses the new simplified Nonlinear PCA learning rule (30) or Lambert's algorithm (14). This also shows that the Amari - Cichocki - Yang learning rule is in fact maximizing the sum of negentropies of the elements y_i of the output vector, eq. (15), while at the same time keeping the outputs uncorrelated.

4.3 Relationship with the EASI algorithm

A well-known, good gradient algorithm for BSS without prewhitening is the EASI algorithm. EASI is introduced and justified as an adaptive signal processing algorithm by Cardoso and Laheld in [2], but it can as well be used as a learning algorithm of a nonlinear PCA type network. The general update formula for the separating matrix $\mathbf{B}$ is in EASI

$$\Delta\mathbf{B} = \mu_k[\mathbf{I} - \mathbf{y}\mathbf{y}^T - \mathbf{g}(\mathbf{y})\mathbf{y}^T + \mathbf{y}\mathbf{g}(\mathbf{y}^T)]\mathbf{B} \tag{32}$$

The EASI algorithm is derived in [2] using the so-called relative gradient concept, which is the same as the natural gradient mentioned earlier.

We can again derive a closely related algorithm from the nonlinear PCA rule (4) using the general update rule (23). Recalling that $\mathbf{v} = \mathbf{V}\mathbf{x}$, $\mathbf{y} = \mathbf{B}\mathbf{x} = \mathbf{W}^T\mathbf{V}\mathbf{x}$, and that $\mathbf{W}^T\mathbf{W} = \mathbf{W}\mathbf{W}^T = \mathbf{I}$, we get for the first term in the square brackets in (23) from (4)

$$\begin{aligned}
\Delta\mathbf{W}^T\mathbf{V} &= \mathbf{g}(\mathbf{y})[\mathbf{x}^T\mathbf{V}^T - \mathbf{g}(\mathbf{y}^T)\mathbf{W}^T]\mathbf{V} \\
&= \mathbf{g}(\mathbf{y})[\mathbf{x}^T\mathbf{V}^T\mathbf{W}\mathbf{W}^T\mathbf{V} - \mathbf{g}(\mathbf{y}^T)\mathbf{B}] \\
&= \mathbf{g}(\mathbf{y})[\mathbf{y}^T - \mathbf{g}(\mathbf{y}^T)]\mathbf{B}
\end{aligned} \tag{33}$$

Using again the simple whitening algorithm (24) yields for the second term $\mathbf{W}^T\Delta\mathbf{V}$ the expression (25). Inserting these results into (23) yields for the total separating matrix $\mathbf{B}$ the learning rule

$$\Delta\mathbf{B} = \mu_k[\mathbf{I} - \mathbf{y}\mathbf{y}^T + \mathbf{g}(\mathbf{y})\mathbf{y}^T - \mathbf{g}(\mathbf{y})\mathbf{g}(\mathbf{y}^T)]\mathbf{B} \tag{34}$$

A comparison with the EASI algorithm (32) shows that the derived algorithm (34) differs only slightly from EASI (the sign of the nonlinear part $\mathbf{g}(\mathbf{y})\mathbf{y}^T - \mathbf{g}(\mathbf{y})\mathbf{g}(\mathbf{y}^T)$ is not important in practice). In [2], the EASI algorithm is derived in a somewhat different way by making rather heavy but sensible approximations.

In a few experiments made with the algorithm (34), it converged somewhat slower than the EASI algorithm (32). After initial convergence the two algorithms behaved quite similarly, giving almost exactly the same results. An explanation is that when the components of the output vector $\mathbf{y}$ become independent, both $\mathrm{E}\{\mathbf{g}(\mathbf{y})\mathbf{g}(\mathbf{y}^T)\}$ and $\mathrm{E}\{\mathbf{y}\mathbf{g}(\mathbf{y}^T)\}$ tend to identity matrices $\mathbf{I}$ multiplied by a constant.

5 Least-squares methods for nonlinear PCA criterion

We have recently shown [16, 18] that the nonlinear PCA criterion (8) can be efficiently minimized using approximative recursive least-squares (RLS) techniques. Generally, RLS algorithms converge clearly faster than their stochastic gradient counterparts, and achieve a good final accuracy at the expense of somewhat higher computational load [10]. These advantages result from the automatic determination of the learning parameter from the input data so that it becomes roughly optimal.

The basic symmetric algorithm, adapted for the BSS problem using prewhitened data vectors $\mathbf{v}(t)$, is [16, 18]

$$\begin{aligned}
\mathbf{z}(t) &= \mathbf{g}(\mathbf{W}^T(t-1)\mathbf{v}(t)) = \mathbf{g}(\mathbf{y}(t)), \\
\mathbf{h}(t) &= \mathbf{P}(t-1)\mathbf{z}(t), \\
\mathbf{m}(t) &= \mathbf{h}(t)/(\beta + \mathbf{z}^T(t)\mathbf{h}(t)), \\
\mathbf{P}(t) &= \frac{1}{\beta}\mathrm{Tri}\left[\mathbf{P}(t-1) - \mathbf{m}(t)\mathbf{h}^T(t)\right], \\
\mathbf{e}(t) &= \mathbf{v}(t) - \mathbf{W}(t-1)\mathbf{z}(t), \\
\mathbf{W}(t) &= \mathbf{W}(t-1) + \mathbf{e}(t)\mathbf{m}^T(t). \qquad (35)
\end{aligned}$$

The forgetting constant $0 < \beta \leq 1$ should be close to unity. The notation Tri means that only the upper triangular part of the argument is computed and its transpose is copied to the lower triangular part, making thus the matrix $\mathbf{P}(t)$ symmetric. The initial values $\mathbf{W}(0)$ and $\mathbf{P}(0)$ can be chosen to $m \times m$ unit matrices.

The RLS algorithm (35) can be regarded either as a neural network learning algorithm or adaptive signal processing algorithm. It is a modified version of the PAST algorithm introduced by Yang for the standard linear PCA in [24]. We have presented the respective sequential RLS algorithm as well as experimental results in [16, 18]. These algorithms have a close connection to nonlinear PCA subspace rule [16, 18]. A batch version of the RLS approach is briefly introduced in [16].

6 Removal of prewhitening

BSS approaches using prewhitening perform usually well, but may suffer from a serious loss of accuracy if some of the source signals $s_i(t)$ are weak or the mixing matrix $\mathbf{A}$ is ill-conditioned [2]. In order to avoid these drawbacks, some researchers [2, 5] have favored so-called equivariant algorithms. These algorithms have the property that the overall system matrix $\mathbf{C}(t) = \mathbf{B}(t)\mathbf{A}(t)$ describing the mixing and demixing process depends only on its previous value $\mathbf{C}(t-1)$ and on the output vector $\mathbf{y}(t) = \mathbf{B}(t)\mathbf{x}(t) = \mathbf{C}(t)\mathbf{s}(t)$. Examples of equivariant algorithms are the natural gradient algorithm (22), EASI algorithm (32) and the algorithm (34). However, if noise is present the equivariance property is lost.

We can modify the nonlinear PCA criterion (2) so that prewhitening is not needed as follows. The Bussgang cost function $\mathrm{E}\{||\mathbf{y} - \mathbf{g}(\mathbf{y})||^2\}$ which is used widely in blind equalization [10] is defined without any signal model for $\mathbf{y}$. Assume now that $\mathbf{y}$ obeys the linear model $\mathbf{y} = \mathbf{Bx}$. We can then express the Bussgang cost with respect to the mixture vector $\mathbf{x}$. Assuming that $\mathbf{B}$ is invertible so that $\mathbf{x} = \mathbf{B}^{-1}\mathbf{y}$ we get

$$\begin{aligned} \mathrm{E}\{||\mathbf{y} \;&-\; \mathbf{g}(\mathbf{y})||^2\} = \mathrm{E}\{||\mathbf{B}\mathbf{B}^{-1}(\mathbf{Bx} - \mathbf{g}(\mathbf{y}))||^2\} \\ &= \; \mathrm{E}\{||\mathbf{x} - \mathbf{B}^{-1}\mathbf{g}(\mathbf{y})||^2_{\mathbf{R}}\} = J_{\mathbf{R}}. \end{aligned} \tag{36}$$

Clearly, $J_{\mathbf{R}}$ is a weighted least-squares cost with the weighting matrix $\mathbf{R} = \mathbf{B}^T\mathbf{B}$. In the special case where $\mathbf{B}$ is orthogonal this reduces to the standard nonlinear PCA cost (2) (with $\mathbf{B} = \mathbf{W}^T$) because $\mathbf{B}^T = \mathbf{B}^{-1}$ and $\mathbf{R} = \mathbf{I}$ when $\mathbf{B}$ is orthogonal.

In the following, we derive a stochastic gradient algorithm for minimizing the cost (36) with respect to the weight matrix $\mathbf{B}$. Denote by $F'(\mathbf{B})$ the gradient matrix whose element kl is the derivative $\frac{\partial F(\mathbf{B})}{\partial b_{kl}}$, where b_{kl} is the respective element of $\mathbf{B}$. We get

$$J' = (\mathbf{M}^T\mathbf{M})' = (\mathbf{M}^T)'\mathbf{M} + \mathbf{M}^T\mathbf{M}' \tag{37}$$

where

$$\mathbf{M} = \mathbf{Bx} - \mathbf{g}(\mathbf{Bx}). \tag{38}$$

After some straightforward computations this approach leads to the gradient algorithm

$$\Delta\mathbf{B} = [\mathbf{y} - \mathbf{g}(\mathbf{y}) - \mathbf{y}\mathbf{g}'(\mathbf{y}) + \mathbf{g}(\mathbf{y})\mathbf{g}'(\mathbf{y})]\mathbf{x}^T. \tag{39}$$

Applying the natural gradient approach [25] where essentially the right hand side of (39) is multiplied by $\mathbf{B}^T\mathbf{B}$ yields the final algorithm

$$\begin{aligned}\Delta\mathbf{B} &= [\mathbf{y} - \mathbf{g}(\mathbf{y}) - \mathbf{y}\mathbf{g}'(\mathbf{y}) + \mathbf{g}(\mathbf{y})\mathbf{g}'(\mathbf{y})]\mathbf{y}^T\mathbf{B} \\ &= [\mathbf{y} - \mathbf{g}(\mathbf{y})][\mathbf{1} - \mathbf{g}'(\mathbf{y})]\mathbf{y}^T\mathbf{B}\end{aligned} \tag{40}$$

Here $\mathbf{1}$ denotes the vector which has ones as its all elements, and the vector products of the form $\mathbf{h}(\mathbf{y})\mathbf{f}(\mathbf{y})$ are understood as Hadamard products or elementwise products. The algorithm (40) has again the desirable equivariance property. If $\mathbf{1} - \mathbf{g}'(\mathbf{y})$ is approximated by $\mathbf{1}$, (40) becomes the natural gradient version of Lambert's algorithm (14).

It is possible to construct an equivariant algorithm from the RLS learning rule (35), too. From the last row of (35), $\Delta\mathbf{W}^T = \mathbf{m}(t)\mathbf{e}^T(t)$. Inserting $\mathbf{e}(t)$ yields after similar manipulations as in subsection 4.2

$$\Delta\mathbf{W}^T\mathbf{V} = \mathbf{m}(t)[\mathbf{y} - \mathbf{z}]^T\mathbf{B}. \tag{41}$$

Note that this term already contains the learning parameter. If the latter term $\mathbf{W}^T\Delta\mathbf{V}$ in (23) is now taken the same $[\mathbf{I} - \mathbf{y}\mathbf{y}^T]\mathbf{B}$ as in (25), one gets an equvariant algorithm for the total separating matrix $\mathbf{B}$. Regrettably, the resulting algorithm suffers from serious stability problems. If $\mathbf{y}$ is normalized so that its squared norm equals to m (stemming from the condition $\mathrm{tr}\mathrm{E}\{\mathbf{y}\mathbf{y}^T\} = \mathrm{E}\{\mathbf{y}^T\mathbf{y}\} = m$), the stability is clearly improved but still the algorithm eventually tends to diverge after a fairly fast initial convergence towards a separating solution.

Therefore, it is better to realize the constraint condition $\mathrm{E}\{\mathbf{y}\mathbf{y}^T\} = \mathbf{I}$ in another way instead of adding the term $\mu_k[\mathbf{I} - \mathbf{y}\mathbf{y}^T]\mathbf{B}$ to (41). Because $\mathrm{E}\{\mathbf{y}\mathbf{y}^T\} = \mathbf{B}\mathbf{R}_{xx}\mathbf{B}^T$ where $\mathbf{R}_{xx}$ is the correlation matrix of input vectors $\mathbf{x}$, the row vectors $\mathbf{b}_i$ of $\mathbf{B}$ must satisfy the generalized orthonormality condition $\mathbf{b}_i^T\mathbf{R}_{xx}\mathbf{b}_j$ $= 1$ if $i = j$ and $= 0$ if $i \neq j$. These conditions can be realized by using the generalized Gram-Schmidt algorithm

$$
\begin{aligned}
\mathbf{b}_1 &= \mathbf{b}_1/(\mathbf{b}_1^T\mathbf{R}_{xx}\mathbf{b}_1)^{1/2}; \\
\text{for} \quad & i = 2, \ldots, m \\
\mathbf{b}_i' &= \mathbf{b}_i - \sum_{j=1}^{i-1}(\mathbf{b}_i^T\mathbf{R}_{xx}\mathbf{b}_j)\mathbf{b}_j \\
\mathbf{b}_i &= \mathbf{b}_i'/[(\mathbf{b}')_i^T\mathbf{R}_{xx}\mathbf{b}_i']^{1/2}.
\end{aligned}
\tag{42}
$$

This "postwhitening" approach performs well, and in our experiments it did not suffer from any stability problems. Furthermore, because $\mathbf{b}_i^T\mathbf{R}_{xx}\mathbf{b}_j = \mathrm{E}\{y_iy_j\}$, the orthonormalization (42) can be realized by adaptively estimating the elements of the covariance matrix $\mathrm{E}\{\mathbf{y}\mathbf{y}^T\}$ of the output vector $\mathbf{y}$.

7 Experimental results

Simulation results for the RLS algorithm (35) both in stationary and nonstationary cases can be found in [16, 18]. They show that RLS type algorithms converge clearly faster than the nonlinear PCA subspace rule (3) or (4) and achieve a good final accuracy.

In experiments with the algorithm (40), we used four sub-Gaussian source signals of 512 samples: a ramp, a sinusoid, a binary signal, and uniformly distributed white noise. These sources were mixed linearly using a 4×4 randomly chosen mixing matrix. The input data are shown in Figure 1 . In each experiment, the samples were iterated 10 times, so that the total number of iterations were 5120. No whitening was used. Results of a simulation are depicted in Figure 2. The algorithm (40) achieves a good accuracy. However, it does not always separate all the sources for the same reason as the algorithm (14), explained in subsection 4.1.

In another experiment, we compared the performance of the algorithm in which the update $\Delta\mathbf{B}$ of the total separating matrix $\mathbf{B}$ was computed using (41) and the rows of $\mathbf{B}$ were then normalized using (42), with the EASI algorithm (32). The same four sub-Gaussian sources as in the first experiment were used. Figure 3 shows the evolution of a standard error measure [25]

$$
E_2 = \sum_{i=1}^{m}(\sum_{j=1}^{m}\frac{p_{ij}^2}{max_k p_{ik}^2} - 1) + \sum_{j=1}^{m}(\sum_{i=1}^{m}\frac{p_{ij}^2}{max_k p_{kj}^2} - 1). \tag{43}
$$

for both the algorithms. The error index E_2 describes how far the matrix $\mathbf{P} = (p_{ij}) = \mathbf{BA}$ defining the input-output mapping is from the ideal result, permutation matrix. The larger the value of E_2, the poorer is the performance, the minimum being zero.

Clearly, the RLS type postwhitening algorithm converges faster than EASI. In other experiments made by us, the results were qualitatively similar with a few exceptions. More detailed experimental comparisons of various neural learning algorithms proposed for Independent Component Analysis and blind source separation can be found in [7].

8 Conclusions

The starting point of this paper is a nonlinear PCA criterion and the respective stochastic gradient algorithm. We have examined the connections of this approach to some other well-known criteria and algorithms, including in particular the Bussgang criterion used in blind deconvolution and the EASI algorithm which is used in blind source separation. We have also derived several new algorithms based on the nonlinear PCA criterion which need not prewhitening.

References

[1] A. Bell and T. Sejnowski, "An information-maximisation approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129-1159, 1995.

[2] J.-F. Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. on Signal Processing*, vol. 44, pp. 3017-3030, December 1996.

[3] J.-F. Cardoso, "Infomax and maximum likelihood for blind source separation," *IEEE Signal Processing Letters*, vol. 4, pp. 112-114, April 1997.

[4] J.-F. Cardoso, "Entropic contrasts for source separation," to appear as Chapter 2 in S. Haykin (ed.), *Adaptive Unsupervised Learning*, 1998.

[5] A. Cichocki and R. Unbehauen, "Robust neural networks with on-line learning for blind identification and blind separation of sources," *IEEE Trans. on Circuits and Systems-1*, vol. 43, pp. 894-906, November 1996.

[6] P. Comon, "Independent component analysis - a new concept?," *Signal Processing*, vol. 36, pp. 287-314, 1994.

[7] X. Giannakopoulos, "Comparison of adaptive independent component analysis algorithms," Dipl.Eng thesis made for EPFL, Switzerland, at Helsinki Univ. of Technology, Finland, 58 p. Available at `http://www.cis.hut.fi/~xgiannak/`.

[8] M. Girolami and C. Fyfe, "Stochastic ICA contrast maximisation using Oja's nonlinear PCA algorithm," submitted to *Int. J. Neural Systems*, July 1996.

[9] S. Haykin (Ed.), *Blind Deconvolution.* Prentice-Hall, 1994.

[10] S. Haykin, *Adaptive Filter Theory, 3rd ed.* Prentice-Hall, 1996.

[11] C. Jutten and J. Herault, "Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, no. 1, pp. 1-10, July 1991.

[12] 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.

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

[14] J. Karhunen, "Neural approaches to independent component analysis and source separation," in *Proc. 4th European Symp. on Artificial Neural Networks (ESANN'96)*, Bruges, Belgium, April 1996, pp. 249-266.

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

[16] J. Karhunen and P. Pajunen, "Blind source separation using least-squares type adaptive algorithms," in *Proc. 1997 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'97)*, Munich, Germany, April 1997, pp. 3361-3364.

[17] J. Karhunen, E. Oja, L. Wang, R. Vigario, and J. Joutsensalo, "A class of neural networks for independent component analysis," *IEEE Trans. on Neural Networks*, vol. 8, pp. 486-504, May 1997.

[18] J. Karhunen and P. Pajunen, "Blind source separation and tracking using nonlinear PCA criterion: a least-squares approach," in *Proc. 1997 Int. Conf. on Neural Networks (ICNN'97)*, Houston, Texas, June 1997, pp. 2147-2152.

[19] R. Lambert, *Multichannel Blind Deconvolution: FIR Matrix Algebra and Separation of Multipath Mixtures.* Ph.D. dissertation, Univ. of Southern California, Dept. of Electrical Eng., May 1996.

[20] E. Moreau and O. Macchi, "High-order contrasts for self-adaptive source separation," *Int. J. of Adaptive Control and Signal Processing*, vol. 10, pp. 19-46, 1996.

[21] E. Oja, H. Ogawa, and J. Wangviwattana, "Learning in nonlinear constrained Hebbian networks". In T. Kohonen et al. (Eds.), *Artificial Neural Networks*, North-Holland, Amsterdam, 1991, pp. 385-390.

[22] E. Oja, "The Nonlinear PCA learning rule and signal separation – mathematical analysis," *Neurocomputing,* vol. 17, pp. 25-45, September 1997.

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

[24] B. Yang, "Projection approximation subspace tracking," *IEEE Trans. on Signal Processing*, vol. 43, pp. 95-107, January 1995.

[25] H. Yang and S.-I. Amari, "Adaptive online learning algorithms for blind separation: maximum entropy and minimum mutual information," *Neural Computation*, vol. 9, pp. 1457-1482, 1997.

Biographies of the authors

Juha Karhunen received the Doctor of Technology degree from the Department of Technical Physics of Helsinki University of Technology in 1984. In 1994, he became Docent of statistical signal analysis. Since 1976, he has been with the Laboratory of Computer and Information Science at Helsinki University of Technology, Finland, where he is currently Senior Research Fellow of the Academy of Finland. His current research interests include neural networks, unsupervised learning, and their applications to signal processing. He has published a number of conference and journal papers on these topics, and given invited talks in several international conferences. Dr. Karhunen is a member of IEEE, INNS, and Pattern Recognition Society of Finland.

Erkki Oja received his Dr.Tech. degree (with distinction) from Helsinki University of Technology, Finland, in 1977. He is Professor of Computer Science at the Laboratory of Computer and Information Science, Helsinki University of Technology. Dr. Oja is the author of a number of journal papers and book chapters on pattern recognition, computer vision, and neural computing, and the book "Subspace Methods of Pattern Recognition", which has been translated into Chinese and Japanese. His present research interests are in the study of subspace, PCA, and self-organizing networks, and applying artificial neural networks to computer vision and signal processing. Dr. Oja has served in the scientific and organization committees of a number of recent conferences. He is senior member of the IEEE, member of the Finnish Academy of Sciences, past chairman of the Finnish Pattern Recognition Society, member of the Governing Board of the International Association of Pattern Recognition (IAPR), IAPR Fellow, vice president of the European Neural Network Society (ENNS), and member of the Neural Networks Technical Committee of the IEEE. He is member of the editorial boards of the journals "Neural Networks", "Neural Computation", "Pattern Recognition Letters", and "IEEE Transactions on Neural Networks".

Petteri Pajunen was born in Helsinki, Finland in 1969. He received his M.Sc. degree from the Helsinki University of Technology in 1995. He is currently a Ph.D. candidate in the Laboratory of Computer and Information Science of the Helsinki University of Technology. His research interests include unsupervised learning, independent component analysis, and signal processing applications of neural networks.

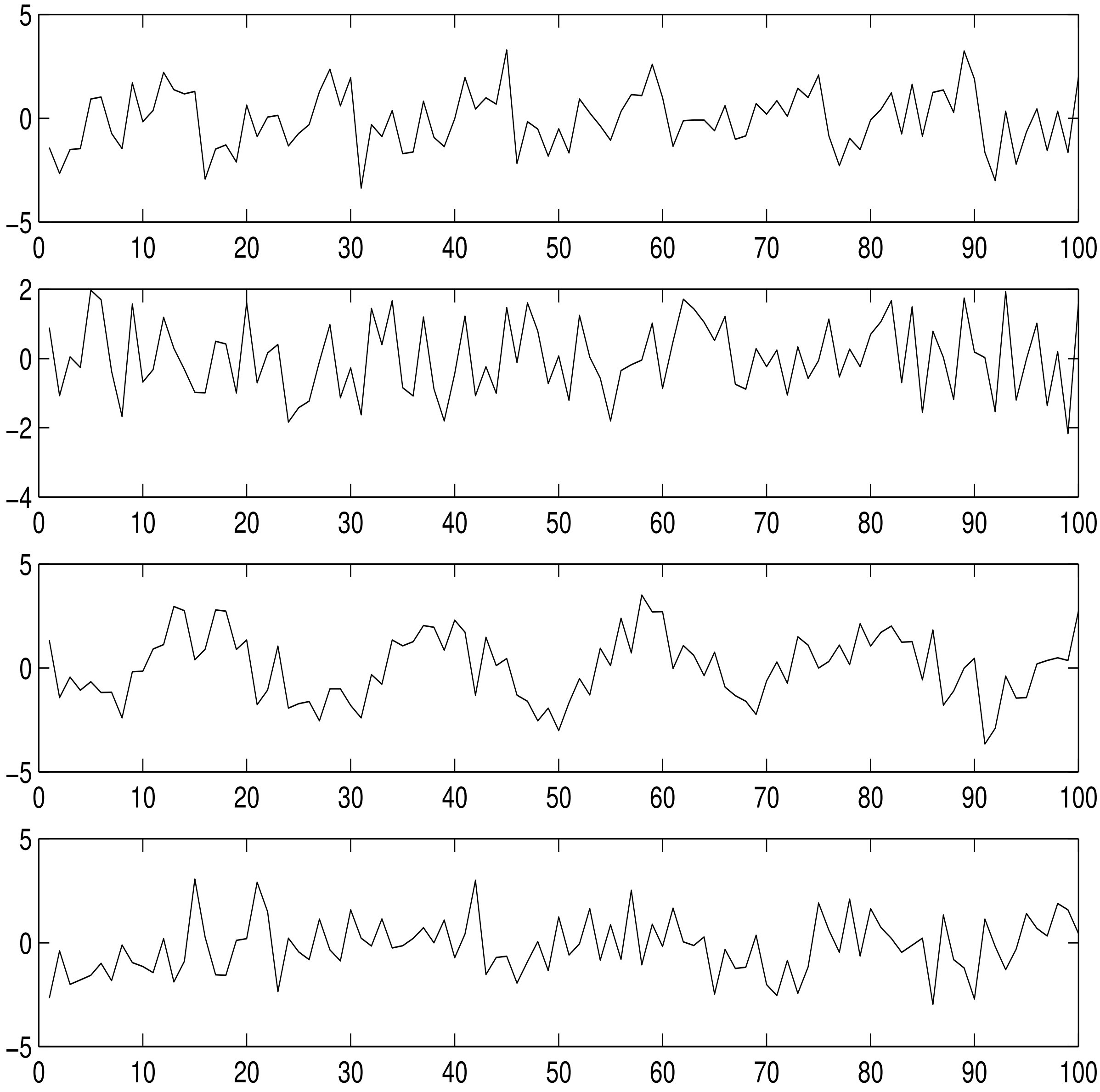

Figure 1: Mixtures of 4 sub-Gaussian sources.

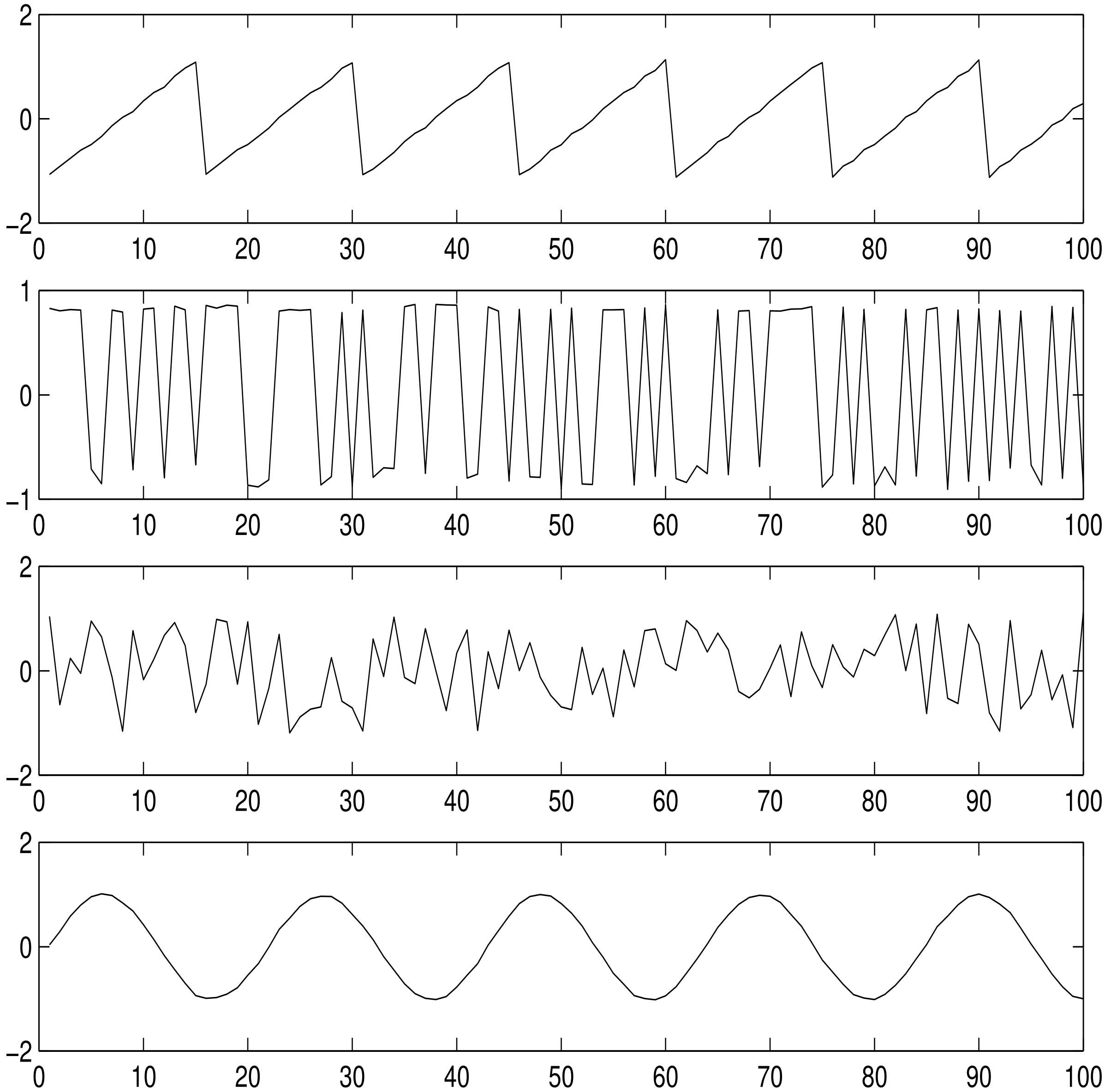

Figure 2: Separated signals.

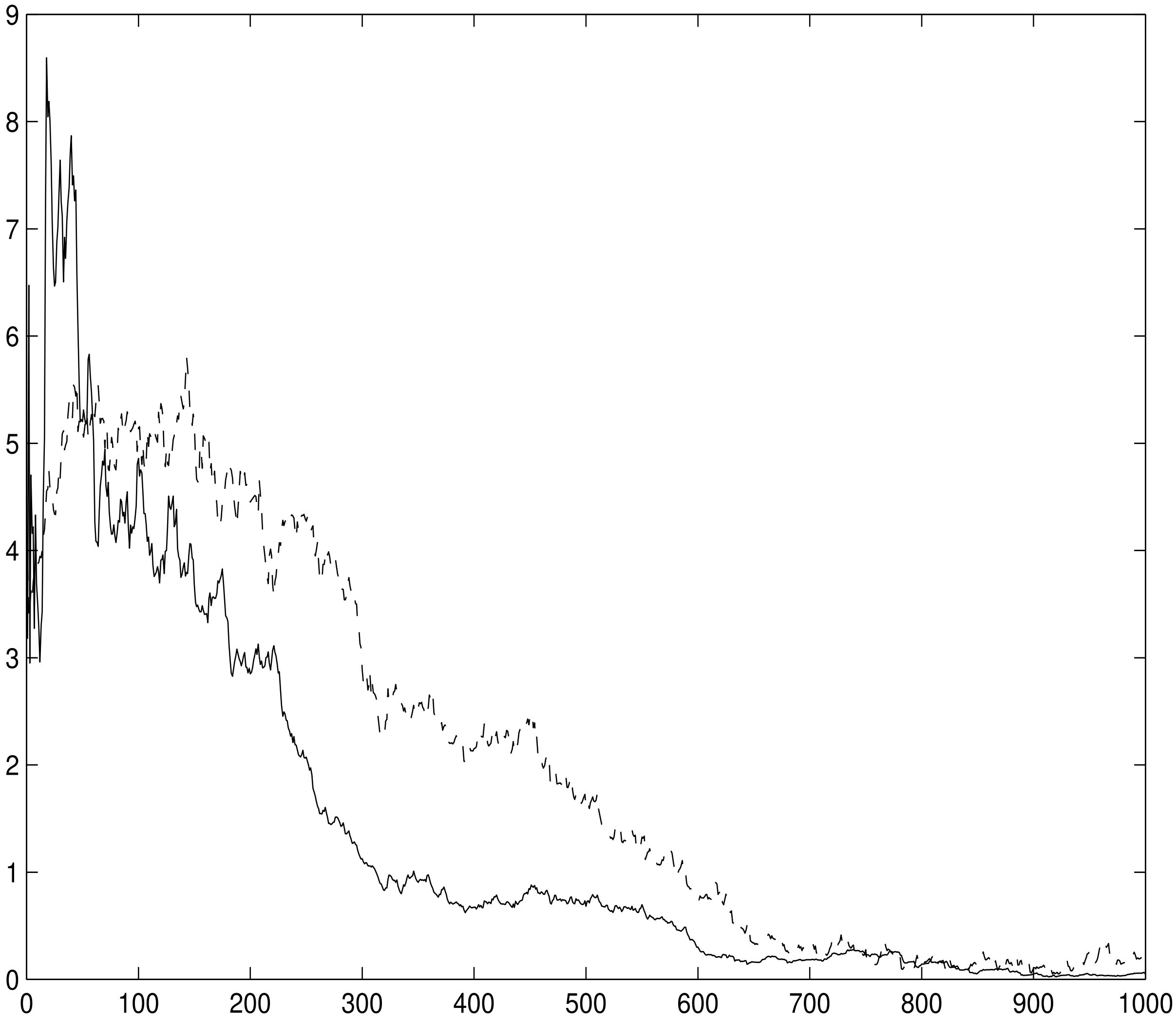

Figure 3: Crosstalk error for the EASI algorithm (dashed line) and RLS algorithm with postwhitening (solid line).