An Unified Perspective of Blind Source Separation Adaptive Algorithms

Sergio Cruces
TSC Group, University of Seville
Camino de los Descubrimientos
41092-Seville, Spain
sergio@viento.us.es

Andrzej Cichocki
Laboratory for Open Information Systems
Brain Science Institute, RIKEN
Wako-Shi, Saitama 351-01, JAPAN
cia@brain.riken.go.jp

Luis Castedo
Dpto. Electrónica y Sistemas
Universidad de La Coruña
Campus de Elviña,
15071-La Coruña, Spain
luis@sol.des.fi.udc.es

Abstract
Blind Source Separation (BSS) is an important topic in signal processing and neural networks where a lot of adaptive algorithms have been recently proposed. These algorithms, however, have been developed from many different points of view. This paper shows that most algorithms can be derived from a single guiding principle: the inversion of a nonlinear correlation matrix. The inversion is carried out in an iterative way using quasi Newton optimization methods and allows us to derive a specific normalized step size that brings high convergence speed and numerical stability. Finally, necessary and sufficient asymptotic stability conditions for the method to converge are given.

1 Introduction
Blind Source Separation (BSS) is a fundamental problem in signal processing that arises in a large number of applications such as array processing, multiuser communications and voice restoration. It is typically formulated as follows. Assume that an array of sensors provides a vector of $N$ observed signals $\mathbf{x}(n) = [x_1(n), \ldots, x_N(n)]^T$ that are mixtures of $N$ random processes $\mathbf{s}(n) = [s_1(n), \ldots, s_N(n)]^T$ termed sources. The exact probability density function of the sources is unknown: it is only assumed that they are zero-mean, non-Gaussian distributed and statistically independent. We will restrict ourselves to the case where the observations are instantaneous linear mixtures of the sources, i.e.,

$$\mathbf{x}(n) = \mathbf{Hs}(n)$$

where $\mathbf{H}$ is a $N \times N$ invertible matrix that represents the mixing system.

To recover the sources, the observations are processed by a Multiple Input Multiple Output (MIMO) system to produce the outputs

$$\mathbf{y}(n) = \mathbf{Ws}(n)$$
where $W$ is the $N \times N$ separating matrix. We will denote $G = WH$ as the matrix that represents the overall mixing and separation system. The aim in BSS is to find or estimate $W$ to retrieve the original sources. This is accomplished when each output corresponds to a single and different source, i.e., when $G$ is the product of a diagonal and a permutation matrix.

BSS is strongly related to neural networks because, due to the Darmois-Skitovich theorem [8], sources are recovered if and only if the outputs of the separating system are statistically independent. Moreover, it has been shown [19] that information transfer in a single layer neural network is maximized when the outputs are statistically independent. Therefore, adaptive algorithms for BSS are valid unsupervised learning rules that seek statistical independence between the neuron outputs and maximize the information transfer.

Since the pioneering work of Jutten and Herault [16] a lot of efficient and robust adaptive algorithms for BSS have been proposed and their properties investigated. These algorithms have been developed from many different points of view such as contrast functions [13, 10], maximum likelihood estimation [4, 3], information transfer maximization [6], Kullback-Leibler minimization using the natural gradient approach [9, 3] and non-linear Principal Component Analysis (PCA) [17]. However, no attempt has been done yet to present them under a unified perspective. The main contribution of the present paper is to demonstrate that most of the above algorithms can be derived from a single guiding principle: the inversion of a nonlinear correlation matrix. The inversion is carried out in an iterative way using quasi Newton optimization methods and allows us to derive a specific normalized step size that supplies high convergence speed and numerical stability to the algorithms.

The paper is structured as follows. Section 2 shows how the BSS separation can be viewed as the inversion of a nonlinear correlation matrix and Section 3 presents two iterative algorithms to perform this inversion. Section 4 interprets some existing BSS algorithms as iterative inversion algorithms. Section 5 derives the asymptotic stability conditions for the iterative inversion algorithms and Section 6 is devoted to the conclusions.

2 BSS as a Matrix Inversion Problem

Several authors [4, 3] have recently interpreted the BSS problem as the estimation of $H^{-1}$ from the observed data $X(n)$. It can be shown [3] that the maximum likelihood estimate $W_\lambda$ corresponds to the minimum of the loss function $L(W) = -E[\log p_X(x|W)]$, being $p_X(x|W)$ the probability density function of $X$ (the random process that generates $x$) conditioned to $W$.

In order to arrive at a more practical form of this loss function, it is assumed that the sources have a hypotheic joint probability density function (pdf) $p_S(s)$ that factors as

$$p_S(s) = \prod_{i=1}^{N} p_i(s_i)$$

where $p_i(s_i)$ is the pdf of $s_i$. This pdf does not need to be equal to the true pdf of the
sources: it is only considered to derive a practical algorithm. Then,
\[ p_S(Wx) = \frac{p_X(x|W)}{|W|} \] (4)
where \(|W|\) is the determinant of \(W\), and the loss function \(L(W)\) can be rewritten as
\[ L(W) = -E[\log(|W|p_S(Wx))] \] (5)

Other authors also arrive at this same optimization problem from different perspectives such as minimizing the Kullback-Leibler distance between the probability density functions of the output and the sources [9] or maximizing the information transfer in a single layer neural network [6].

A necessary condition for the maximum likelihood estimate is [3]
\[ \left( \frac{\partial L(W)}{\partial W} \right)^T_{W=W_*} = 0 \Rightarrow \] (6)
\[ (R_{x\psi} - W^{-1})_{W=W_*} = 0 \Rightarrow \] (7)
\[ W_s = R_{\psi\psi}^{-1} \] (8)
where \( \psi(\cdot) = [-\frac{d\log p_S}{dx_1}, \ldots, -\frac{d\log p_S}{dx_N}]^T \) is a componentwise nonlinear function and \( R_{x\psi} = E[x\psi^T(y)] \) is the cross correlation matrix between the input \( x \) and a nonlinear function of the outputs \( \psi(y) \). Note that this matrix is the maximum likelihood estimator of \( H \) when \( W = W_s \). The condition (8) is of paramount importance for us because it suggests that iterative methods that calculate the inverse of \( R_{x\psi} \) are valid algorithms for BSS.

Note also that BSS is accomplished when a nonlinear autocorrelation matrix of the outputs becomes diagonal. This is clearly seen rewriting condition (8) as
\[ W_sR_{x\psi} = R_{\psi\psi} = I \] (9)

Now the maximum likelihood estimation \( W_s \) corresponds to the separating system that diagonalizes \( R_{\psi\psi} = E[\psi(y)^T(y)] \) or, equivalently, its transposed version \( R_{\psi\psi} = E[\psi(y)y^T] \).

It is interesting to note that when the sources are Gaussian distributed the nonlinear function of the output \( \psi(y) \) reduces to \( \psi(y) = y \) and the nonlinear correlation matrix \( R_{\psi\psi} \) reduces to the conventional autocorrelation matrix \( R_{yy} = E[yy^T] \). Therefore, minimization of the loss function \( L(W) \) results in a simple decorrelation of the output vector \( y \). However, it is well known that decorrelation is not an adequate criterion for blind source separation (neither for unsupervised learning in neural networks). The hypothesized pdf \( p_S(s) \) should correspond to non-Gaussian distributions as stated in the Darmois-Skitovich theorem [8].

In this paper we generalize the above idea and propose as a guiding principle for BSS the diagonalization of the more general nonlinear correlation matrix \( R_{fg} = E[f(y)g^T(y)] \) where \( f(y) \) and \( g(y) \) represent two nonlinear functions of the outputs. The diagonalization of \( R_{fg} \) means that it is equal to the identity at the optimum separation matrix \( W_s \), i.e.,
\[ R_{fg}(W_s) = I \] (10)
At a first glance, it may appear that this generalization does not make sense since only the selection \( f(y) = y \) and \( g(y) = \psi(y) \) provides the maximum likelihood estimator of \( H^{-1} \). However, as pointed out in [4], the consideration of other estimating functions, with two non-linearities, may have preferred global properties and be more robust to Gaussian noise.

The diagonalization of \( R_{fg} \) can also be viewed as a matrix inversion: if we define a new non-linear function of the outputs \( F(y) = W^{-1}f(y) \) and a new nonlinear autocorrelation matrix

\[
R_{Fg}(W) = E[F(y)g^T(y)] = W^{-1}R_{fg}(W)
\]

and

\[
R_{fg}(W_s) = I \Rightarrow W_s = R_{Fg}^{-1}(W_s)
\]

Solving the BSS problem is, thus, equivalent to inverting \( R_{Fg} \).

### 3 Iterative Inversion Algorithms

The matrix \( R_{Fg}(W_s) \) cannot be inverted directly since we do not have access to \( F(y) \) and \( g(y) \) at the optimum separating solution \( W_s \). Instead, we propose to solve the following system of nonlinear equations

\[
\mathcal{F}(W^{-1}) = R_{Fg} - W^{-1} = 0
\]

using a quasi Newton method termed chord method that, in our case, exhibits better convergence properties than conventional Newton-Raphson methods [18]. The resulting recursion when applying the chord method with respect to \( W^{-1} \) is

\[
W^{-1}_{(n+1)} = W^{-1}_{(n)} - \mu \mathcal{F}(W^{-1}_{(n)})B^{(n)}
\]

where the inverse of the derivative matrix, \( \frac{\partial \mathcal{F}}{\partial W^{-1}} \), is approximated by a matrix \( B^{(n)} \) that must satisfy certain conditions described in [18]. In our case we have chosen the approximation \( \left( \frac{\partial \mathcal{F}}{\partial W^{-1}} \right)^{-1} \approx -I = B^{(n)} \) which is the result of assuming that \( R_{Fg} \) hardly depends on \( W \). This is because we can interpret \( R_{Fg} \) as a rough estimate of the mixing system \( H \) as seen from eq. (12). Having in mind this approximation we arrive at the following algorithm

\[
W^{-1}_{(n+1)} = (1 - \mu)W^{-1}_{(n)} + \mu R^{(n)}_{fg}
\]

where \( R^{(n)}_{fg} = R_{fg}(W = W^{(n)}) \). We will designate this algorithm as the first Iterative Inversion (II1) algorithm. It is interesting to note that \( W^{-1}_{(n)} \) is an estimate of the spatial correlation matrix \( R_{Fg} \) with an exponential window. This is consistent with our interpretation of \( R_{Fg} \) as an estimate of the mixing system \( H \).

Next, let us find an iterative algorithm where the adaptation is carried out with respect to \( W^{(n)} \) rather than \( W^{-1}_{(n)} \). Assuming that \( (I + \mu (R^{(n)}_{fg} - I)) \) is nonsingular, the recursion (16) can be inverted to yield

\[
W^{(n+1)} = \left( I + \mu (R^{(n)}_{fg} - I) \right)^{-1} W^{(n)}
\]
It is desirable in classic optimization methods to restrict the algorithms to work in a domain where $\mathcal{F}(\mathbf{W}^{-1})$ and its derivative are continuous. Assuming that $\mathbf{f}(\cdot)$ and $\mathbf{g}(\cdot)$ are twice differentiable functions, the only discontinuities that may occur are located at the points where the separating matrix $\mathbf{W}$ is singular. Therefore, discontinuities in $\mathcal{F}(\mathbf{W}^{-1})$ can be simply avoided if we constrain the algorithm to work in a connected region where $\mathbf{W}$ is never singular.

A sufficient condition to ensure this is to select $\mu^{(n)}$ in such a way that

$$
\mu^{(n)} < \frac{1}{\|\mathbf{R}_{fg}^{(n)} - \mathbf{I}\|_p}
$$

where $\| \cdot \|_p$ denotes the p-norm of a matrix. This way we prevent $(\mathbf{I} + \mu (\mathbf{R}_{fg}^{(n)} - \mathbf{I}))$ from being singular. As a consequence, we can express the inverse in (17) as the following expansion

$$
(\mathbf{I} + \mu^{(n)} (\mathbf{R}_{fg}^{(n)} - \mathbf{I}))^{-1} = \sum_{i=0}^{\infty} (-\mu^{(n)} (\mathbf{R}_{fg}^{(n)} - \mathbf{I}))^i \\
\approx \mathbf{I} - \mu^{(n)} (\mathbf{R}_{fg}^{(n)} - \mathbf{I})
$$

Finally, substituting (19) into (17) we arrive at another quasi-Newton algorithm for the inversion of $\mathbf{R}_{fg}$

$$
\mathbf{W}^{(n+1)} = \mathbf{W}^{(n)} - \mu^{(n)} (\mathbf{R}_{fg}^{(n)} - \mathbf{I}) \mathbf{W}^{(n)}
$$

We will refer to this new recursion as the second Iterative Inversion (II2) algorithm.

To illustrate the validity of the approximation used to derive the II2 algorithm let us denote $\mathbf{W}^{(n+1)}_1$ and $\mathbf{W}^{(n+1)}_2$ the result obtained after applying the III1 and II2 algorithms, respectively, from the same value $\mathbf{W}^{(n)}$, i.e.,

$$
\mathbf{W}^{(n+1)}_1 = \mathbf{W}^{(n)} (\mathbf{I} + \mu (\mathbf{R}_{fg}^{(n)} - \mathbf{I})) \\
\mathbf{W}^{(n+1)}_2 = \mathbf{W}^{(n)} - \mu^{(n)} (\mathbf{R}_{fg}^{(n)} - \mathbf{I}) \mathbf{W}^{(n)}
$$

If both algorithms were identical, the product $\mathbf{W}^{-1} (n+1) \mathbf{W}^{(n+1)}_2$ would be the identity. This will not be the case since, due to the approximation in (19), algorithms III1 and II2 are different. However, it is straightforward to show that following equality holds true

$$
\|\mathbf{W}^{-1} (n+1) \mathbf{W}^{(n+1)}_2 - \mathbf{I}\|_p = \mu^{(n)} \|\mathbf{R}_{fg}^{(n)} - \mathbf{I}\|_p^2
$$

This result means that both algorithms perform almost the same when $\mu^{(n)}$ is small or $\mathbf{R}_{fg}^{(n)}$ is very close to be diagonal, i.e., the algorithms are close to convergence.

It should be noted that the II2 algorithm is the extended batch version of the family of equivariant adaptive algorithms proposed originally by Cichocki et al. in [12, 11] for blind source separation. Also, many existing algorithms for BSS can be derived from the II2 algorithm as explained in section 4.
3.1 Practical Considerations

The iterative inversion algorithms derived in the previous section make use of the non-linear correlation matrix $R_{fg}^{(n)}$ which is a statistical average. This matrix is usually not known in practical situations and should be estimated from the incoming data. There are two possibilities: on-line adaptations in which the correlation matrix $R_{fg}^{(n)}$ is replaced by its single sample estimate, i.e.,

$$R_{fg}^{(n)} \approx f(y(n))g^T(y(n))$$  \hspace{1cm} (23)

and batch adaptations where, if we assume that the stationarity and ergodicity properties hold on a block of $L$ samples, we can replace in (20) the statistical average by the temporal average

$$R_{fg}^{(n)} = \frac{1}{L} \sum_{k=0}^{L-1} f(y(k))g^T(y(k))$$ \hspace{1cm} (24)

Another major practical issue is the selection of the algorithm step-size. We are interested in choosing the step-size so as to ensure both numerical stability and high convergence speed. Recall that to derive the II/2 algorithm we needed to ensure the condition (18). Taking into account that $\|R_{fg}^{(n)} - I\|_p \leq 1 + \|R_{fg}^{(n)}\|_p$ we propose to estimate the learning step as

$$\mu^{(n)} = \frac{\mu_1}{1 + \mu_1\|R_{fg}^{(n)}\|_p} < \frac{1}{1 + \|R_{fg}^{(n)}\|_p}$$  \hspace{1cm} (25)

where $\mu_1 < 1$ is an arbitrary positive constant. In practice we will choose $\mu_1$ close to unity since we are interested in the algorithm to converge as fast as possible. Although other norms are perfectly valid, in batch adaptations we will typically set $p = 1$ or $p = \infty$ since these norms result in the simplest form of the step-sizes.

In the on-line case, however, it is more practical to use $p = 2$ since

$$\|R_{fg}^{(n)}\|_2 = \|g^T(y(n)) f(y(n))\|$$

The resulting practical step-size is thus

$$\mu^{(n)} = \frac{\mu_1}{1 + \mu_1\|g^T(y(n)) f(y(n))\|}$$  \hspace{1cm} (26)

Note that now $\mu_1 \ll 1$ because we are only using single sample estimates of $R_{fg}^{(n)}$ that introduce a lot of misadjustment noise into the algorithm.

4 Relation with Existing Algorithms

Several existing algorithms for BSS can be derived from the II/2 algorithm (20): the decorrelation algorithm of Almeida [2], the natural gradient algorithm for BSS [3], the non-linear PCA algorithm [17] (under the requirement of having an orthogonal separation matrix) and the normalized and generalized EASI algorithms [10, 17]. In fact, all these models can be shown to be consistent with (20) when one or more linear/non-linear correlation matrices are diagonalized. Below, we will show these relations for the on-line case.
**Decorrelation algorithm**: Setting \( f(y) = g(y) = y \) the adaptation rule (20) reduces to the decorrelation algorithm proposed by Almeida [2] (see also [12]). Using the normalized step size (26) the resulting algorithm is

\[
W^{(n+1)} = W^{(n)} - \mu_1 \frac{yy^T - I}{1 + \mu_1 |y^Ty|} W^{(n)}
\]  

(27)

**Amari’s natural gradient algorithm for BSS**: When only one of the two functions \( f(y) \) or \( g(y) \) is non-linear the II2 algorithm converts into

\[
W^{(n+1)} = W^{(n)} - \mu_1 \frac{f(y) y^T - I}{1 + \mu_1 |y^T f(y)|} W^{(n)}
\]  

(28)

or in its permuted form. Both are normalized versions of the Natural Gradient algorithms for BSS developed by Amari [3] and independently by Cardoso and Laheld [10].

**Generalized EASI algorithms**:

The family of generalized EASI algorithms was proposed by Karhunen et al. in [17] as an extension of the family of EASI algorithms derived by Cardoso and Laheld in [10]. We can derive its normalized version from the II2 method following a procedure similar to that described by Cardoso and Laheld in their article. Let us split the separation stage in two blocks connected in series, i.e., \( W = W_b W_a \). The first matrix \( W_a \) will be selected to diagonalize the symmetric matrix \( R_{yy} \) and will be adapted according to algorithm (27). On the other hand, the second matrix \( W_b \) will be selected to diagonalize \( R_{fg} \) and towards this aim the on-line version of the algorithm (20) can be used. However, in order to merge (20) and (28) into a single recursion for the overall separating system \( W \) we need that both adaptations be orthogonal at a first order. Since the matrix \( yy^T - I \) is always symmetric, to orthogonalize the second adaptation at a first order we can replace \( f(y)g^T(y) - I \) by its projection onto the space of skew-symmetric matrices and the following algorithm results

\[
W_b^{(n+1)} = \left( I - \mu_1 \frac{f(y) g^T(y) - g(y) f^T(y)}{1 + \mu_1 |f^T(y) g(y)|} \right) W_b^{(n)}
\]

Combining both adaptation rules as \( W^{(n+1)} = W_b^{(n+1)} W_a^{(n+1)} \) and keeping only the terms of the first order expansion in \( \eta \) (which is a valid approximation for \( \eta \) small enough) we arrive at

\[
W^{(n+1)} = \left( I - \mu_1 \frac{yy^T - I}{1 + \mu_1 |y^T y|} \right. \\
\left. - \mu_1 \frac{f(y) g^T(y) - g(y) f^T(y)}{1 + \mu_1 |f^T(y) g(y)|} \right) W^{(n)}
\]

which is a normalized version of the family of generalized EASI algorithms [17, 10].
Non-linear PCA algorithm:

Imposing the constraint that the separation matrix \( W^{(n)} \) be orthogonal, we can derive the non-linear PCA algorithm developed by Oja and Karhunen [20, 17] from the II2 method. First, let us redefine the nonlinear correlation matrix as

\[
R_{fg}^{(n)} = E[f(y)g^T(y)] + I
\]

where \( f(y) \) is an odd function and \( g(y) = f(y) - y \). Substituting now the matrix \( R_{fg}^{(n)} \) by its stochastic approximation in the II2 algorithm (20) we arrive at

\[
W^{(n+1)} = W^{(n)} - \mu \left( R_{fg}^{(n)} - I \right) W^{(n)}
\]

(29)

Taking into account that \( W^{(n)} \) is an orthogonal matrix, i.e.,
\[ W^T W = I, \]
we can rewrite (29) as

\[
W^{(n+1)} = W^{(n)} + \mu f(y) \left( x^T - f^T(y) W^{(n)} \right)
\]

which is Karhunen-Oja non-linear PCA algorithm [20, 17].

5 Stability analysis

In this section we perform a stability analysis of the II2 algorithm for real sources. We will provide the necessary and sufficient conditions that the nonlinearities must satisfy in order to ensure that the separating solutions are stable points.

We will adopt the following notation: \( f_i = [f(s)]_i, f'_i = \frac{\partial f(s)}{\partial s_i} |_{s_i}, g_i = [g(s)]_i, g'_i = \frac{\partial g(s)}{\partial s_i} |_{s_i}. \) Also, we will assume that the sources \( s \) are properly scaled and the functions \( f(s) \) and \( g(s) \) are properly chosen so the conditions for the separation solution be an stationary point for the algorithm, i.e.,

\[
E[f_i g_i] = 1 \quad \text{and} \quad E[f_i g_j] = 0
\]

(30)

hold for all \( i \neq j \).

Theorem 1 Assuming that

1. \( s \) is the vector of real and zero mean sources.
2. The functions \( f(\cdot) \) and \( g(\cdot) \) act component-wise on the elements of the argument vector and are at least twice differentiable in the domain of support.
3. The step-size \( \mu \) is small enough.

the II2 algorithm with a constant step size given by

\[
W^{(n+1)} = W^{(n)} - \mu \left( R_{fg}^{(n)} - I \right) W^{(n)}
\]

(31)
exhibits a asymptotically stable point at the separation solution if and only if the following conditions are satisfied

\[ E[f_is_i] = 0 \quad \text{or} \quad E[g_j] = 0 \quad (32) \]
\[ E[g_js_i] = 0 \quad \text{or} \quad E[f_j] = 0 \quad (33) \]
\[ E[f_is_ig_i] + E[f_is_ig_i] > 0 \quad (34) \]
\[ E[f_iE[s_ig_j] + E[f_j]E[s_ig_i] > 0 \quad (35) \]
\[ E[f_iE[f_j]E[s_ig_i]E[s_ig_j] > E[g_iE[g_j]E[s_i]E[s_j] \quad (36) \]

for all \( i, j \neq \).

The proof of this theorem is presented in appendix A.

From this theorem we can obtain the following corollary, first presented in [5].

**Corollary 1** When one of the two functions is linear, i.e., \( g(y) = y \) and assuming that the sources are properly scaled so that

\[ E[f_is_i] = 1 \quad (37) \]

for all \( i, j \neq \), the asymptotic stability conditions of the II2 algorithm reduce to the following three

\[ E[f_is_i^2] + 1 > 0 \quad (38) \]
\[ E[f_iE[s_i^2] + E[f_j]E[s_j^2] > 0 \quad (39) \]
\[ E[f_iE[f_j]E[s_i^2]E[s_j^2] > 1 \quad (40) \]

for all \( i, j \).

**Proof:** Note that when \( g(y) = y \) the constraint \( E[f_ig_j] = 0 \) is always true since by the independence and zero mean assumptions for the sources \( E[f_ig_j] = E[f_is_j] = E[g_is_j] = 0, \forall i \neq j \). The conditions (38–40) are straightforwardly obtained from (34–36) taking into account that \( E[g_i] = E[s_i] = 0, E[g_is_i] = E[s_i] = 0 \) and \( g_i' = 1 \) for all \( i \). □

If the sources have even probability density functions and the nonlinearities are strictly monotonically increasing odd functions, conditions (32–35) are always satisfied. These assumptions, however, do not guarantee that condition (36) be true in which case the following corollary is useful to stabilize the algorithm.

**Corollary 2** If

\[ E[f_i]E[f_j]E[s_ig_i]E[s_ig_j] \neq E[g_iE[g_j]E[s_is_i]E[s_js_j] \quad (41) \]

an interchange of the nonlinear functions ensures the asymptotical stability of the II2 algorithm.

**Proof:** Since the sources probability density functions are even and the nonlinearities strictly odd, \( E[f(s)] = 0 \) and \( E[g(s)] = 0 \). On the other hand, when the nonlinearities are strictly monotonically increasing odd functions \( E[f_is_ig_i] > 0 \) and \( E[f_is_ig_i] \) and thus condition (34) is true. For the same reasons, \( E[f_is_i^2] > 0 \) and \( E[s_ig_i] > 0 \) \( \forall i \) and condition
(34) is also true. Finally, condition (34), when the equality does not hold, is not as critical since it can be forced by just permuting the order of the functions $f(\cdot)$ and $g(\cdot)$.

This fact has been also observed in [21] for the Jutten-Herault algorithm and later in [3] when $g(y) = y$. Our results, however, are valid for the general case of two non-linear functions and for the II2 algorithm.

Finally,

**Corollary 3** The II1 algorithm

\[
(W(n+1))^{-1} = (W(n))^{-1} \left( I + \mu(R_{fg}^{(n)} - I) \right)
\]  

(42)

has the same asymptotical stability conditions given by (32-36) as the II2 algorithm.

**Proof:** As demonstrated in section 3 both algorithms behave the same when the step size is small or the algorithm is close to the convergence. □

### 6 Conclusions

A lot of adaptive algorithms exist to solve the BSS problem that, however, have been developed from many different points of view. In this paper we show that most of them can be viewed under a unified perspective: as quasi Newton iterative algorithms that perform a diagonalization of a non linear autocorrelation matrix of the outputs. This approach allows us to derive a normalized step size that results in fast convergent and numerically stable algorithms. We also obtain the necessary and sufficient asymptotic stability conditions for the method to converge towards the separation solutions.

### Acknowledgments

Part of this work is the result of a stay at the Brain Science Institute (RIKEN), supported by the Frontier Research Program of Japan. This work has also been supported by the Grants: CICYT (TIC96-0500-C10-08), CICYT (TIC96-0500-C10-02) and Xunta de Galicia (XUGA 10502A96).

### A Stability Analysis of the II2 Algorithm

In this appendix we will obtain the asymptotic stability conditions (32-36) for the II2 algorithm under the assumptions of Theorem 1. Since the algorithm is equivariant [10, 3] we can multiply the algorithm by the mixing matrix $H$ to obtain the following recursion for the overall transfer matrix $G$.

\[
G^{(n+1)} = G^{(n)} - \mu \left( R_{fg}^{(n)} - I \right) G^{(n)}
\]  

(43)

Let $C(Gs) = f(Gs)g^T(Gs) - I$. The first condition of the theorem (30) is necessary to allow the separation solution $G_s = I$ to be an equilibrium point of the algorithm, i.e., to ensure that $E[C(G_s)] = 0$. 

10
The next step is the study of the algorithm asymptotic stability at this point using the Ordinary Differential Equation (ODE) method [7]. Let us define the algorithm mean field as \( M(G) = E[C(Gs)G] \) and an arbitrary small perturbation of \( G \), as \( \epsilon \). Then,
\[
M(I + \epsilon) = E[C(s + \epsilon s)(I + \epsilon)] = E[C(s + \epsilon s)] + o(\epsilon)
\]
In order to find the linear approximation to the mean field in terms of \( \epsilon \), we will replace both functions \( f(s + \epsilon s) \) and \( g(s + \epsilon s) \) by its first order Taylor expansion at the solution
\[
\begin{align*}
f(s + \epsilon s) &= f(s) + f'(s)\epsilon s + o(\epsilon) \\
g(s + \epsilon s) &= g(s) + g'(s)\epsilon s + o(\epsilon)
\end{align*}
\]
where \( f'(s) = \frac{\partial f(s)}{\partial s} \) and \( g'(s) = \frac{\partial g(s)}{\partial s} \) are diagonal matrices since \( f(\cdot) \) and \( g(\cdot) \) act componentwise. Substituting (45) in (44) and using \( E[C(Gs)] = 0 \) at the stationary point,
\[
M(I + \epsilon) = E[f'(s)sg^T(s)] + E[f(s)s^T\epsilon^T(g'(s))^T] + o(\epsilon)
\]
Adopting the notation \( f_i = [f(s)]_i \), \( f'_i = [f'(s)]_i \), the terms of the mean field are
\[
M_{ij}(I + \epsilon) = \sum_k E[f'_i s_k g_j] \epsilon_{ik} + \sum_k E[f_i s_k g'_j] \epsilon_{jk} + o(\epsilon)
\]
Taking into account that the sources are zero-mean, statistically independent and satisfy
\[
\begin{align*}
E[f'_i s_i] &= 0 \quad \text{or} \quad E[g_j] = 0 \\
E[g'_i s_i] &= 0 \quad \text{or} \quad E[f_j] = 0
\end{align*}
\]
we arrive at
\[
M_{ij}(I + \epsilon) = \begin{cases} 
(E[f'_i s_i g_i] + E[f_i s_i g'_i])\epsilon_{ii} + o(\epsilon) & \text{if } i = j \\
E[f'_i] E[s_i g_j] \epsilon_{ij} + E[g'_j] E[s_i f_i] \epsilon_{ji} + o(\epsilon) & \text{if } i \neq j
\end{cases}
\]
Let us consider now the following vector parametrization of the ODE associated to the algorithm (43)
\[
\theta_{t+1} = \theta_t - \mu \psi(\theta_t, z_t)
\]
where \( z_t \) is the state vector. Denoting \( \theta^* \) to the stationary point of the algorithm that gives the separation solution, the mean vector field of the algorithm at this point is \( \mathcal{M}(\theta^*, z) = E[\psi(\theta^*, z)] \). It is well known that the asymptotic stability condition for the algorithm at \( \theta^* \) is that all the eigenvalues of the derivative of the mean vector field, at that point, have all positive real parts.

Taking a convenient parametrization for the separation matrix \( G \) in the vector \( \theta \) as
\[
\theta = \left[ G_{ij}, G_{ji}, \ldots, G_{ii}, \ldots \right]^T
\]
and similarly for the mean field \( M(G) \) in \( \mathcal{M}(\theta) \), it is found that the gradient of the main field at the separation \( \frac{\partial M(G)}{\partial \theta} \bigg|_{\theta=\theta^*} \) takes a very simple form: it is block-diagonal, formed only by diagonal blocks of size one and two. This fact, that has been found before in
several blind source separation algorithms (see for example [10, 20]), will facilitate to
found closed analytical expressions for asymptotic stability conditions.

One of the conditions for asymptotic stability is extracted from the one-size blocks of the diagonal

\[
\frac{\partial M_{ii}(I + \epsilon)}{\partial \epsilon_{ii}} = E[f_i^s(i)g_i] + E[f_i^s(i)g_i^s] > 0
\]  

(53)

The other blocks of the derivative matrix of size two are of the form

\[
\left( \begin{array}{cc}
\frac{\partial M_{ii}(I+\epsilon)}{\partial \epsilon_{ii}} & \frac{\partial M_{ij}(I+\epsilon)}{\partial \epsilon_{ij}} \\
\frac{\partial M_{ji}(I+\epsilon)}{\partial \epsilon_{ij}} & \frac{\partial M_{jj}(I+\epsilon)}{\partial \epsilon_{jj}}
\end{array} \right)
= \left( \begin{array}{cc}
E[f_i^sE[s_jg_i] & E[g_i^sE[s_jf_i]] \\
E[g_i^sE[s_jf_i]] & E[f_i^sE[s_jg_i]]
\end{array} \right)
\]  

(54)

for all \( i \neq j \). The eigenvalues \( \lambda \) of these block matrices are the roots of their characteristic polynomial equation. In order to check if the real part of the eigenvalues is positive, we can apply the necessary and sufficient Hurwitz stability conditions [1], obtaining the latest necessary and sufficient conditions for the asymptotic stability of the proposed algorithm

\[
E[f_i^sE[s_jg_i] + E[f_i^sE[s_jg_i]] > 0
\]  

(55)

\[
E[f_i^sE[f_j^sE[s_jg_i]]E[s_jg_j] > E[g_i^sE[g_j^sE[s_i f_i]]E[s_jf_j]]
\]  

(56)

for all \( i \neq j \)

References


