On a New Blind Signal Extraction Algorithm: Different Criteria and Stability Analysis

Sergio A. Cruces-Alvarez, *Associate Member, IEEE*, Andrzej Cichocki, *Member, IEEE*, and Shun-Ichi Amari, *Fellow, IEEE*

***Abstract*—In this letter, we consider the problem of simultaneous blind signal extraction of arbitrary group sources from a rather large number of observations. Amari proposed a gradient algorithm that optimizes the maximum-likelihood (ML) criteria on the Stiefel manifold and solves the problem when the approximate (or hypothetical) densities of the desired signals are *a priori* known. This letter shows how to extend this result to other contrast functions that do not require explicit knowledge of the sources densities. We also present the algorithm necessary and sufficient local stability conditions, providing useful bounds for the learning step size.**

***Index Terms*—Blind source separation, contrast functions, independent component analysis, simultaneous blind signal extraction.**

I. Introduction

BLIND SIGNAL separation (BSS) is the problem of recovering mutually independent unobserved signals (sources) from their linear mixture. Although this problem has recently attracted a lot of interest because of its wide number of applications in diverse fields, BSS can be very computationally demanding if the number of source signals is large (say, of order 100 or more). In particular, this is the case in biomedical signal processing applications such as electroencephalographic/magnetoencephalographic data processing, where the number of sensors can be larger than 120 and where it is desired to extract only some "interesting" sources. Fortunately, sequential blind signal extraction (BSE) overcomes somewhat this difficulty. The BSE problem considers the case where only a small subset of sources has to be recovered from a large number of sensor signals.

The combined use of BSE and deflation to solve the BSS problem was originally proposed in [3] and later further explored in several papers (e.g., [4]). However, the main limitation of existing BSE algorithms is that most of them can only recover the sources sequentially, one by one, in order to avoid the possibility of obtaining the sources replicated at the outputs. In this letter, we will present a straightforward technique that allows the extension of several of the classical criteria for blind source separation and extraction to the case of the simultaneous blind extraction of an arbitrary subgroup of P ($1 \leq P \leq N$, where N means the total number of sources) interesting sources.

Let us consider the standard linear mixing model of N unknown statistically independent source signals drawn from a random vector process $\mathbf{S}(t) = [S_1(t), \ldots, S_N(t)]^T$ of zero mean and normalized covariance $\mathrm{Cov}(\mathbf{S}) = E[\mathbf{S}(t)\mathbf{S}^T(t)] = \mathbf{I}_N$, where $\mathbf{I}_N$ denotes the identity matrix of rank N. These signals are linearly combined by the memoryless system described by a mixing matrix $\mathbf{A}$ to give the observations

$$\mathbf{X}(t) = \mathbf{A}\mathbf{S}(t). \tag{1}$$

Without loss of generality, we assume that the unknown mixing matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ is orthogonal, and since the mixture is instantaneous, we can drop the time reference when referring to the random variables of the considered processes. Note, that the orthogonality of the mixing matrix $(\mathbf{A}\mathbf{A}^T = \mathbf{I}_N)$ can be always enforced by simply performing prewhitening on the original observations. For noisy data, the robust prewhitening or orthogonalization can be employed.

In order to extract $P \leq N$ sources, the observations will be further processed by a $P \times N$ semi-orthogonal separating matrix $\mathbf{U}$, satisfying $\mathbf{U}\mathbf{U}^T = \mathbf{I}_P$, which yields to the outputs vector (or estimated sources)

$$\mathbf{Y} = \mathbf{U}\mathbf{X} = \mathbf{G}\mathbf{S} \tag{2}$$

where $\mathbf{G} = \mathbf{U}\mathbf{A}$ will denote the semi-orthogonal $P \times N$ global transfer matrix from the sources to the outputs. The semi-orthogonality of the global transfer system will be important for preserving the spatial decorrelation of the outputs vector, since $\mathrm{Cov}(\mathbf{Y}) = \mathbf{G}\mathbf{G}^T = \mathbf{I}_P$.

According to the usual notation, random variables and their samples are denoted in capital and lowercase letters, respectively. We will denote by $C_Y^r = \mathrm{Cumulant}(Y : r)$ to the rth-order cumulant of the random variable Y, by $h(Y) = -\int p_Y(y) \log p_Y(y) dy$ to its differential entropy, and by $\delta[\cdot]$ to the discrete Dirac's delta or unit impulse signal.

II. Criteria for Simultaneous BSE

Let us define a functional $\psi(\cdot)$ that maps each density of a normalized random variable (n.r.v.) Y_i (of zero mean and unit variance) to a real index $\psi(Y_i)$ that satisfies the following *properties*.

Manuscript received November 13, 2001; revised May 24, 2002. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Giovanni Ramponi.

S. A. Cruces-Alvarez is with Area de Teoría de la Señal y Comunicaciones, Escuela de Ingenieros, Universidad de Sevilla, 41092-Sevilla, Spain (e-mail: sergio@us.es).

A. Cichocki is with the Laboratory for Advanced Brain Signal Processing, Brain Science Institute, Riken, Japan, and also with the Department of Electrical Engineering, Warsaw University of Technology, Poland (e-mail: cia@bsp.brain.riken.go.jp).

S.-I. Amari is with the Brain-Style Information Systems Research Group, Brain Science Institute, Riken, Japan (e-mail: amari@brain.riken.go.jp).

Publisher Item Identifier 10.1109/LSP.2002.803025.

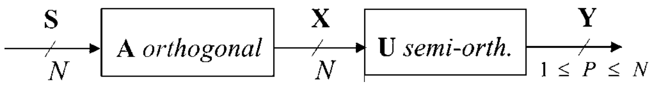

Fig. 1. Considered signal model for simultaneous blind source extraction.

1) $\psi(\cdot) \geq 0$, and the minimum value of the index ($\psi(Y_i) = 0$) is obtained when $p_{Y_i} = p_G$, i.e., when the n.r.v. follows a Gaussian distribution.
2) $\psi(\cdot)$ is convex (strictly convex) with respect to the linear combinations of the independent sources (of the independent sources for which $\psi(S_j) \neq 0$), in such a way that, if $Y_i = \sum_{j=1}^{N} G_{ij} S_j$, then

$$\psi(Y_i) \leq \sum_{j=1}^{N} |G_{ij}|^2 \psi(S_j) \tag{3}$$

where G_{ij} are the elements of the semi-orthogonal matrix $\mathbf{G}$ and $S_j, j = 1, \ldots, N$, are independent and normalized random variables.

These properties have been defined in [5], and they are close to those given in [6] and [7] when the idea of contrast functions was independently defined. Since then, many functionals that satisfy the previous properties have been proposed in the literature, such us, among others, the minimum entropy (ME)-based index [5], [7]

$$\psi_{\mathrm{ME}}(Y_i) = \frac{1}{2}\log(2\pi e) - h(Y_i) \tag{4}$$

the maximum-likelihood (ML)-based index [1, Ch. 3]

$$\psi_{\mathrm{ML}}(Y_i) = \text{constant} + E[\log p_{S_i}(y_i)] \tag{5}$$

and the cumulant-based indexes [3], [4], [6], [8]

$$\psi_C(Y_i) = \sum_{r>2} \omega'_r |C^r_{Y_i}|^{\alpha_r} \tag{6}$$

where $\alpha_r \geq 1$ (typically $\alpha_r = 1$ or 2) and $\omega'_r = \omega_r/(r\alpha_r)$ are scaled or normalized nonnegative weighting factors.

From Property 2), the blind extraction of one of the non-Gaussian sources is obtained, solving the following constrained maximization problem

$$\max_{\mathbf{U}} \psi(Y_i) \text{ subject to } \mathrm{Cov}(Y_i) = 1 \tag{7}$$

whereas it is well known [6] that the blind source separation of the whole set of sources is obtained, maximizing

$$\max_{\mathbf{U}} \sum_{i=1}^{N} \psi(Y_i) \text{ subject to } \mathrm{Cov}(\mathbf{Y}) = \mathbf{I}_N. \tag{8}$$

The next theorem fills the theoretical gap between both previous approaches.

Theorem 1: Given a set of positive constants $d_1 > d_2 > \cdots > d_P$ and a functional $\psi(\cdot)$ that satisfies Properties 1)–2), if the sources can be ordered by decreasing the value of this functional as

$$\psi(S_1) \geq \cdots \geq \psi(S_P) > \psi(S_{P+1}) \geq \cdots \geq \psi(S_N) \tag{9}$$

and if $\psi(S_P) \neq 0$, then, the following objective function

$$\Psi(\mathbf{Y}) = \sum_{i=1}^{P} d_i \psi(Y_i) \text{ subject to } \mathrm{Cov}(\mathbf{Y}) = \mathbf{I}_P \tag{10}$$

will be a contrast function whose global maxima correspond to the extraction of the first P sources from the mixture. If, additionally, $\psi(S_1) > \cdots > \psi(S_P)$, then the global maximum is unique and corresponds to the ordered extraction of the first P sources of the mixture, i.e., at this maximum $\mathbf{Y} = [S_1, \ldots, S_P]^T$.

Proof: From Property 2), we have that

$$\sum_{i=1}^{P} d_i \psi(Y_i) \leq \sum_{j=1}^{N} \psi(S_j) \sum_{i=1}^{P} d_i |G_{ij}|^2 \tag{11}$$

$$= \mathrm{trace}\{\mathbf{G}\boldsymbol{\Lambda}\mathbf{G}^T\mathbf{D}\} \tag{12}$$

where

$$\boldsymbol{\Lambda} = \mathrm{diag}(\psi(S_1), \ldots, \psi(S_N)) \text{ and } \mathbf{D} = \mathrm{diag}(d_1, \ldots, d_P)$$

are diagonal matrices. But the decorrelation constraint for the outputs $(\mathrm{Cov}(\mathbf{Y}) = \mathbf{I}_P)$ is tantamount to the semi-orthogonality of the global transfer matrix $\mathbf{G}$. From the application of the Poincaré's separation theorem of matrix algebra, and according to the sources ordering (9), the eigenvalues $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_P$ of $\mathbf{G}\boldsymbol{\Lambda}\mathbf{G}^T$ satisfy $\forall i = 1, \ldots, P$

$$\psi(S_{N-P+i}) \leq \sigma_i \leq \psi(S_i). \tag{13}$$

Thus, taking into account (13) and the majorization of the diagonal elements of the matrix $\mathbf{G}\boldsymbol{\Lambda}\mathbf{G}^T$ by its eigenvalues, the maximum of (12) subject to the semi-orthogonality of $\mathbf{G}$ is

$$\max_{\mathbf{G}\mathbf{G}^T = \mathbf{I}_P} \mathrm{trace}\{\mathbf{G}\boldsymbol{\Lambda}\mathbf{G}^T\mathbf{D}\} = \sum_{i=1}^{P} d_i \psi(S_i) \tag{14}$$

and, from the strict convexity of $\psi(\cdot)$ whenever $\psi(\cdot) \neq 0$, if $\psi(S_1) > \cdots > \psi(S_P)$, the necessary and sufficient condition for the equality between (11) and (14) is that $\mathbf{G} = [\mathbf{I}_P, \mathbf{0}]$, i.e., $\mathbf{G}$ is the ordered extraction matrix of the first P sources. On the other hand, if $\psi(S_1) \geq \cdots \geq \psi(S_P)$ with equality for certain subsets of the first P sources that have a common value or index under $\psi(\cdot)$, the necessary and sufficient condition for the equality between (11) and (14) is that the matrix $\mathbf{G}$ can be reduced to the form $[\mathbf{I}_P, \mathbf{0}]$ by permutations among the rows associated with the sources that share the same index.

III. Extraction Algorithm

A particularly simple and useful method to maximize any chosen contrast function is to use the natural Riemannian gradient ascent in the Stiefel manifold of semi-orthogonal matrices, which is given by

$$\tilde{\nabla}_{\mathbf{U}}\Psi = \nabla_{\mathbf{U}}\Psi - \mathbf{U}(\nabla_{\mathbf{U}}\Psi)^T\mathbf{U}. \tag{15}$$

This leads to the algorithm proposed in [1, Ch. 3] for blind source extraction using the ML approach

$$\mathbf{U}^{(n+1)} = \mathbf{U}^{(n)} - \mu\left(\mathbf{D}\mathbf{R}^{(n)}_{\varphi,x} - \mathbf{R}^{(n)}_{y,\varphi}\mathbf{D}\mathbf{U}^{(n)}\right) \tag{16}$$

TABLE I
PARTIAL ACTIVATIONS FUNCTIONS $\varphi^{(r)}(\cdot)$ ASSOCIATED TO LOW ORDER CUMULANTS

$r=3$	$\rightarrow$	$\varphi^{(3)}(y_i) = -\mathrm{sign}(C^3_{y_i}) \cdot \lvert C^3_{y_i}\rvert^{(\alpha_3-1)} \cdot y_i^2$
$r=4$	$\rightarrow$	$\varphi^{(4)}(y_i) = -\mathrm{sign}(C^4_{y_i}) \cdot \lvert C^4_{y_i}\rvert^{(\alpha_4-1)} \cdot (y_i^3 - 3y_i E[y_i^2])$
$r=5$	$\rightarrow$	$\varphi^{(5)}(y_i) = -\mathrm{sign}(C^5_{y_i}) \cdot \lvert C^5_{y_i}\rvert^{(\alpha_5-1)} \cdot (y_i^4 - 4y_i E[y_i^3] - 6y_i^2 E[y_i^2])$
$r=6$	$\rightarrow$	$\varphi^{(6)}(y_i) = -\mathrm{sign}(C^6_{y_i}) \cdot \lvert C^6_{y_i}\rvert^{(\alpha_6-1)} \cdot (y_i^5 - 5y_i E[y_i^4] - 10y_i^2 E[y_i^3] - 10y_i^3 E[y_i^2] + 30y_i(E[y_i^2])^2)$

where $\mathbf{D} = \mathrm{diag}(d_1,\ldots,d_P)$, $\mathbf{R}^{(n)}_{\varphi,x} = E_t[\boldsymbol{\varphi}(\mathbf{y})\ \mathbf{x}^T]$ is a sample cross correlation matrix, and $\varphi(\mathbf{y}) = [-(d\tilde{\psi}_1(y_1))/(dy_1),\ldots,-(d\tilde{\psi}_P(y_P))/(dy_P)]^T$ is the activation function that depends on $\tilde{\psi}_i(\cdot)$, the stochastic form of the index $\psi(Y_i) = E[\tilde{\psi}_i(Y_i)]$.

The activation functions for the ML-based index $\Psi(\cdot) = \Psi_{\mathrm{ML}}(\cdot)$ can be explicitly computed only when the sources densities are known. However, this is not necessary for the ME-based index $\Psi(\cdot) = \Psi_{\mathrm{ME}}(\cdot)$, and approximations of the true activation functions can be found in [6] and in [1, Ch. 3] using the truncated Edgeword and Gram–Charlier expansion of the marginal probability density functions of the outputs. The most robust approach is for the cumulant-based index, since for this case the general form of activation function can be obtained without approximations, and it is universal in the sense that it does not depend on the density of sources for well-defined indexes (those that are nonzero for all desired sources). The activation function is

$$\varphi(y_i) = \sum_{r>2} \omega_r \varphi^{(r)}(y_i) \tag{17}$$

which is a linear combination of partial activation functions $\varphi^{(r)}(y_i)$ where each one is related with only one rth-order cumulant. The expressions of the partial activation functions are explicitly shown in Table I up to order six, although, in practice, cumulants with order $r>5$ are scarcely used, since their precise estimation requires a large number of samples.

Our objective is to extract the desired sources, i.e., the source signals in which the largest index $\psi(S_i)$ or a contrast value. However, since we use a gradient algorithm, it can be trapped in the local maxima corresponding to other valid extracting solutions. Thus, there is no guarantee that we achieve always the global maximum solution in one single stage of extraction. Fortunately, extensive simulation experiments show that it is usually sufficient to repeat the extraction procedure two or three times with a deflation procedure (see [3] for more details on deflation) to obtain all desired signals, which are, in our case, those with the largest index $\psi(Y_i)$ among all possible estimated sources. An alternative approach is to run the algorithm starting with different initial conditions. The procedure can be stopped when all the sources in the last extraction exhibit small indexes.

IV. STABILITY ANALYSIS AND PRACTICAL IMPLEMENTATION OF THE ALGORITHM

In this section, we will consider an arbitrary nonlinear componentwise activation function $\varphi(\cdot)$, and we will denote it briefly as $\varphi_i = \varphi_i(S_i)$ when acting on the ith extracted source. The theorem and its corollary present the obtained stability results.

Theorem 2: Assuming that the mixing system is orthogonal, the necessary and sufficient local stability conditions of the gradient algorithm in the Stiefel manifold (16) to converge to the extraction solution are, for all $i,j\,|_{i\neq j} = 1,\ldots,P$, given by

$$0 < \mu < \frac{2}{\kappa_i}, \qquad \text{if } P = 1 \tag{18}$$

$$0 < \mu < \min\left\{\frac{2}{\kappa_i+\kappa_j}, \frac{2}{\kappa_i}\right\}, \qquad \text{if } 1 < P < N \tag{19}$$

$$0 < \mu < \frac{2}{\kappa_i+\kappa_j}, \qquad \text{if } P = N \tag{20}$$

where the variables $\kappa_i = d_i\, E[(\partial\varphi_i)/(\partial s_i) - S_i\varphi_i]$ (originally defined in [2]) control the stability of the algorithm.

Proof: Multiplying (16) from the right by the mixing system $\mathbf{A}$, we can study the convergence of iteration in terms of the global transfer system $\mathbf{G}$. In the neighborhood of the extraction solution, we define a global system in terms of the deviation matrix $\epsilon = [\epsilon_L, \epsilon_R]$ as $\mathbf{G} = [\mathbf{I}_P + \epsilon_L, \epsilon_R]$, where ϵ_L is skew-symmetric in order to preserve the first order semi-orthogonality of $\mathbf{G}$. After several manipulations of the iteration, the linearized dynamic of the algorithm around the extraction point is obtained as

$$\epsilon_{ij}^{(n+1)} = (1 - \mu(\kappa_i + \kappa_j))\epsilon_{ij}^{(n)} \tag{21}$$

$$\epsilon_{iq}^{(n+1)} = (1 - \mu\kappa_i)\epsilon_{iq}^{(n)} \tag{22}$$

for $i,j\,|_{i\neq j} = 1,\ldots,P$ and $q = P+1,\ldots,N$. Thus, the necessary and sufficient asymptotic stability condition that enforces ϵ_{ij} and ϵ_{iq} to converge to zero with the run of iterations yields the presented bounds (18)–(20) for the algorithm step size. □

However, it is interesting to observe, that since Amari's algorithm takes the special form of the EASI algorithm [2] in the particular case of $P = N$ and $\mathbf{D} = \mathbf{I}_N$, the condition for the blind separation of all the sources (19) is a simple extension of the local stability condition of the EASI algorithm [2]

$$\kappa_i + \kappa_j > 0, \qquad \text{for all } 1 \le i < j \le N. \tag{23}$$

The results of the following corollary, when applied to Theorem 2, help to express explicitly the bounds for the step size for each of the three considered indexes.

Corollary 1: The factors that control the local stability are

$$\kappa_i = d_i\left(E\left[\left(\frac{\partial \log p_{S_i}}{\partial s_i}\right)^2\right] - 1\right) > 0 \tag{24}$$

for the entropy- and ML-based indexes and

$$\kappa_i = d_i \sum_{r>2} \omega_r |C_{S_i}^r|^{\alpha_r} > 0 \tag{25}$$

for the cumulant-based index.

Proof: We have seen that the stability of the algorithms critically depends on the κ_i factors and on their positiveness. For the entropy- and ML-based contrasts, we can apply the Cramer–Rao lower bound for the estimation of the sources, together with their unit variance constraint and non-Gaussianity, to obtain the result (24). The proof of the second part of the theorem relies on the fact that, for the activation function obtained from the cumulant-based index, $E[(\partial\varphi_i)/(\partial s_i)] = 0$ and, thus, for this special case

$$\kappa_i = -d_i E[S_i \varphi_i] = d_i \sum_{r>2} \omega_r |C_{S_i}^r|^{\alpha_r} \tag{26}$$

which, for any well-defined index $\psi_C(S_i) \neq 0$, is always positive for all $i = 1, \ldots, P$. □

As a consequence of the corollary, if this later index is well defined for the extracted sources, even when the sources densities are unknown, the activation function (17) guarantees the local stability of the algorithm for a sufficient small step size [whose upper bounds are given in (18)–(20)].

V. Simulations

As an exemplary simulation, we consider random mixtures of 100 normalized sources (with zero mean and unit variance); five are asymmetric binary sources with probability mass function $p_S(s) = 0.2\delta[s-2]+0.8\delta[s+0.5]$, and 95 are binary symmetric sources $p_S(s) = 0.5\delta[s+1]+0.5\delta[s-1]$. In order to favor the simultaneous extraction of the asymmetric sources from the mixture, we use an index based on cumulants of odd order (note that this index will vanish for the symmetric sources). We chose cumulants of order 3, i.e., $\omega_r = \delta[r-3]$ and $\alpha_3 = 1$. We set to 5 the number of sources to extract $P = 5$, and we performed 100 random simulations and used the histogram to distinguish the desired sources among those estimated. In each simulation, we ran the simultaneous extraction algorithm one or several times (with deflation in between) until all the asymmetric sources were recovered. The obtained results were that, in 22% of the experiments in which we extracted all the desired sources with just the first run of the algorithm, this quantity increases to 96% of the experiments if a second run is allowed and to 100% after the third run. Thus, we can observe that the used index possesses a high capability of discrimination of asymmetric sources. Similar results have been also obtained for continuous distributions.

VI. Conclusion

In this letter, we have analyzed the problem of the blind simultaneous signal extraction, extending the work of [1, Ch. 3] to other criteria. We obtained activation functions that are universal and, thus, robust when the sources densities are unknown. We also presented the local stability conditions of Amari's extraction algorithm, providing useful bounds for its step size. Finally, via simulations, we have shown that there are situations for which it is possible to find indexes that provide discrimination capability and allow to favor the simultaneous extraction of signals, with desired stochastic properties, from a large number of sensors.

References

[1] S. Amari, A. Cichocki, and H. H. Yang, "Blind signal separation and extraction," in *Unsupervised Adaptive Filtering*, S. Haykin, Ed. New York: Wiley, 2000, vol. I.

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

[3] N. Delfosse and P. Loubaton, "Adaptive blind separation of independent sources: A deflation approach," *Signal Process.*, vol. 45, pp. 59–83, 1995.

[4] A. Hyvarinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Comput.*, vol. 9, pp. 1483–1492, 1997.

[5] P. J. Huber, "Projection pursuit," *Ann. Stat.*, vol. 13, no. 2, pp. 435–475, 1985.

[6] P. Comon, "Independent component analysis, a new concept?," *Signal Process.*, vol. 3, no. 36, pp. 287–314, 1994.

[7] D. Donoho, "On minimum entropy deconvolution," in *Applied Time Series Analysis II*, D. F. Findley, Ed. New York: Academic, 1981.

[8] E. Moreau and O. Macchi, "High-order contrasts for self-adaptive source separation criteria for complex source separation," *Int. J. Adaptive Control Signal Proc.*, vol. 10, pp. 19–46, 1996.

[9] K. E. Hild, II, D. Erdogmus, and J. Príncipe, "Blind source separation using Renyi's mutual information," *IEEE Signal Processing Lett.*, vol. 8, pp. 174–176, June 2001.

[10] R. W. Brockett, "Dynamical systems that sort list, diagonalize matrices, and solve linear programming problems.," *Linear Algebra Appl.*, vol. 146, pp. 79–91, 1991.