Robust blind source separation algorithms using cumulants

Sergio Cruces[a,*], Luis Castedo[b], Andrzej Cichocki[c,1]

[a] *Area de Teoría de la Señal y Comunicaciones, Escuela de Ingenieros, Universidad de Sevilla,*
Camino de los descubrimientos s/n, 41092-Sevilla, Spain
[b] *Dpto. Electrónica y Sistemas, Universidad de La Coruña, Campus de Elviña, 15071-La Coruña, Spain*
[c] *Laboratory for Open Information Systems, Brain Science Institute, RIKEN, Japan*

Abstract

In this paper we propose a new approach to blind separation of independent source signals that, while avoiding the imposition of an orthogonal mixing matrix, is robust with respect to the existence of additive Gaussian noise in the mixture. We demonstrate that, for the wide class of source distributions with certain non-null cumulants and a pre-specified scaling, separation is always a saddle point of a cumulant-based cost function. We propose a quasi-Newton approach for determining this saddle point. This enables us to obtain a family of separation algorithms which, based on higher order statistics, yields unbiased estimates even in the presence of large Gaussian noise and has the interesting property of local isotropic convergence. Another family of algorithms that incorporates second-order statistics loses the former desirable convergence properties but it provides more precise estimates in the absence of noise. Extensive computer simulations confirm robustness and the excellent performance of the resulting algorithms. © 2002 Elsevier Science B.V. All rights reserved.

1. Introduction

When sensors are used to collect information, we typically find that the signals provided by them are linear mixtures of the signals of interest. We then have the

---

* Corresponding author. Tel.: +34-954-487475; fax: +34-954-487373.
  E-mail addresses: sergio@us.es (S. Cruces), luis@sol.des.fi.udc.es (L. Castedo), cia@brain.riken.go.jp (A. Cichocki).
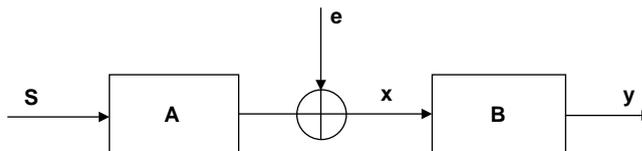 [1] On leave from Warsaw University of Technology, Poland.

Fig. 1. Signal model for blind separation in noise.

problem of recovering the signals of interest, termed sources, from the observed signals. This problem, commonly known as blind source separation (BSS), is aggravated by the fact that in many practical applications we cannot directly observe the sources nor the way that they are mixed. BSS plays an important role in signal processing because it can be solved using a minimum amount of prior information [5,10,14], namely that the mixing system is invertible and that the sources are non-Gaussian and mutually independent. This model robustness makes BSS attractive to a large number of extremely diverse applications such as array processing, multiuser communications, signal restoration and biomedical engineering.

Fig. 1 shows the signal model considered in BSS. Let us denote $\mathbf{s} = [s_1[n], s_2[n], \ldots, s_N[n]]^T$ as the vector of source signals. We will assume that the sources are real, zero mean and that their exact joint probability density function (p.d.f.) is unknown except for the fact that their $(\beta + 1)$-order cumulants $Cum(s_i[n], \ldots, s_i[n])$ are all equal to ± 1. Note that this is not any practical limitation since, in general, we can find a common $\beta > 1$ for which all the sources have non-zero $(\beta + 1)$-order cumulants, as long as the sources are non-Gaussian. Therefore, for the sake of simplicity, we can remove the scaling indeterminacy of the BSS problem by normalizing the modulo of these cumulants to the unity.

Let us next assume that the sensors provide a vector of observed signals $\mathbf{x} = [x_1[n], x_2[n], \ldots, x_N[n]]^T$ that are a memoryless linear combination of the sources, which are also contaminated by the presence of additive Gaussian noise. Thus

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{e}, \tag{1}$$

where $\mathbf{A}$ is the unknown $N \times N$ invertible matrix that represents the mixing system and $\mathbf{e} = [e_1[n], e_2[n], \ldots, e_N[n]]^T$ is the vector of noise components which are assumed in this paper to be Gaussian and statistically independent of the sources.

According to a corollary of the Darmois–Skitovich theorem [8], we can guarantee the identifiability of the mixture, up to a possible scaling and reordering of the sources, if we assume that the source signals are mutually independent and at most one of them is Gaussian. The independence assumption enables us to express the joint p.d.f. of the sources, $p_{\mathbf{S}}(\mathbf{s})$, as the product of the marginal p.d.f. of each source, $p_{S_i}(s_i)$, i.e.,

$$p_{\mathbf{S}}(\mathbf{s}) = \prod_{i=1}^{N} p_{S_i}(s_i). \tag{2}$$

To recover the sources from the observations we use a separating system represented by a $N \times N$ transfer matrix $\mathbf{B}$ to produce the outputs

$$\mathbf{y} = \mathbf{B}\mathbf{x}. \tag{3}$$

The aim in BSS is to select **B** such that each output of the separating system retrieves a single and different original source. The separation is obtained when the overall transfer system $\mathbf{G} = \mathbf{BA}$ is non-mixing or transparent, i.e., it can be reduced to the identity matrix by simply reordering and scaling the sources. For the sake of simplicity in the notation, in this paper we will consider the special reordering that results when the overall transfer system is the identity. This consideration will not affect the contributions of the paper.

Since the pioneering work of Jutten and Herault [19], many efficient and robust adaptive algorithms for BSS have been proposed and their properties investigated. These algorithms have been developed from different points of view such as contrast functions [14,26], maximum likelihood estimation [4,22,29], information transfer maximization [6], Kullback–Leibler divergence minimization using the natural gradient approach [2,9], and non-linear principal component analysis (PCA) [20,28]. Despite their disparity, all these adaptive rules have a common characteristic: the utilization of non-linear functions of their outputs. This non-linearity is required because second-order statistics are not sufficient to solve the BSS problem. However, their convergence properties are extremely dependent on the sources distribution and the non-linearities used. Moreover, these properties are considerably different when there is Gaussian noise in the mixture.

In this work we propose a new family of adaptive algorithms for BSS that only use higher order cumulant functions of the outputs. We demonstrate that, unlike existing algorithms, the local convergence of our approach is isotropic and independent of the sources' distributions as long as these cumulants be non-zero for the sources. Additionally, the utilization of higher order cumulants as non-linearities ensures us that the convergence properties remain unchanged when there is Gaussian noise. However, when only small data sets of observations are available, the use of higher order cumulants results in less accurate estimates. To overcome this limitation we also propose a second family of algorithms that includes second-order statistics information and thus exhibits more accurate convergence.

Although recently, other algorithms such as the JADE algorithm [12], the Extended-Infomax algorithm [17,24], the Fast-ICA algorithm [18] and the Zarzoso–Nandi algorithm [31], have been shown to be able to separate sources from a broad class of distributions; their convergence properties are still dependent on the sources statistics and most of them need the mixing matrix be orthogonal. This is an important point since a mixture with additive noise in the observations is usually a more realistic situation and, trying to impose the orthogonality of the mixing system for this case turns out to be a very sensitive and non-trivial task. This is especially true when the noise is spatially correlated and the number of sensors is not much greater than the number of sources, so that much of the noise subspace cannot be removed. Thus, due to this sensitivity, the algorithms will not be operating under their required theoretical assumptions since a certain lack of orthogonality will exist in the mixing matrix which will favor the existence of a bias in the obtained estimates. As opposed to these approaches, the algorithms we propose here will be robust in this conditions. They skip the previous difficulty because for them the orthogonality of the mixing system is no longer necessary.

This paper is organized as follows. In Section 2 we introduce our cumulant-based approach to BSS which relates the separation solution with the specific saddle points of a given function. In Sections 3 and 4 we present the first family of asymptotically equivariant algorithms that converge to these saddle points and investigate their convergence properties. The incorporation of second-order information is analyzed in Section 5. Section 6 extend the algorithms to the complex case and also to the situation of more sensors than sources. Section 7 presents the results of computer simulations that corroborate the obtained theoretical results and, finally, Section 8 is devoted to the conclusions.

Along the paper we will use the following notation. The operator $diag(\cdot)$ will work according to the MatLab convention, when the operand is a matrix, $diag(\cdot)$ will result in a vector with the matrix's diagonal elements, whereas, when the operand is a vector it will produce the inverse operation forming a diagonal matrix with the vector elements. The operator $vec(\cdot)$ will stack the columns of a square matrix of dimension $N \times N$ into a vector $\mathbf{m} = vec(\mathbf{M})$ of dimension $N^2 \times 1$ with the usual ordering. The trace operator will be denoted by $tr\{\cdot\}$, the Kronecker product by $\otimes$, the Hadamard product by $\odot$ and the α Hadamard power by $(\cdot)^{\odot\alpha}$. We will use $C_{y_i}^{\alpha}$ to denote the α-order cumulant of the output y_i. Analogously, we will denote by $\mathbf{C}_{y,y}^{\alpha,\beta}$ to the matrix whose (i,j) element is given by the cross-cumulant function $C_{y_i,y_j}^{\alpha,\beta} = Cum(\underbrace{y_i,\ldots,y_i}_{\alpha},\underbrace{y_j,\ldots,y_j}_{\beta})$. $\mathbf{S}_y^{\beta}$ will be a short-hand notation to refer to the diagonal matrix containing the signs of the diagonal cumulants $\mathbf{S}_y^{\beta} = diag(sign(diag(\mathbf{C}_{y,y}^{1,\beta})))$. Note that, at the separation, $\mathbf{S}_s^{\beta} = \mathbf{C}_{s,s}^{1,\beta}$ by assumption.

2. Source separation as a saddle point

In this section we will explain the main idea of our approach. It basically consists in the determination of a function for which the BSS solution, regardless of the sources densities, will always be a special kind of critical point whose properties can be later exploited for its localization. As we will show below, the proposed method is radically different from the usual approaches for blind separation because the desired critical point is not maximum or a minimum of the function but a saddle point.

Since the key assumption for the separation of the sources is their mutual independence, the most natural guiding principle for BSS is the minimum mutual information (MMI) principle [14,30], which states that the separation can be always found as the global minimum of the mutual information of the outputs.

In the absence of noise, $\mathbf{y} = \mathbf{Gs}$ and the p.d.f. of the outputs can be expressed in terms of the p.d.f. of the sources as $p_Y(\mathbf{y}) = p_S(\mathbf{y})/|det(\mathbf{G})|$. Thus, mutual information of the sources can be written as

$$I(Y_1,\ldots,Y_N) = \sum_{i=1}^{N} h(y_i) - \log|\det(\mathbf{G})| - h(\mathbf{s}), \tag{4}$$

where $h(\mathbf{s})$ is a constant that represents the differential entropy of $\mathbf{s}$. It is interesting to observe the role of the different terms in this equation. The first term in the right-hand side of (4), the sum of the marginal entropies of the outputs, is a measure of the non-Gaussianity of the distribution. On the other hand, the second term, $-\log|\det(\mathbf{G})|$, normalizes the scaling of the outputs preventing us from reaching the trivial solution $\mathbf{G} = \mathbf{0}$.

The minimization of the mutual information is not trivial because in the first term we need to estimate the p.d.f. of the outputs from a finite set of data. The main approaches to overcome this problem involve the truncation of the Edgeworth [14] or Gram–Charlier [30] expansions of the p.d.f. These truncations, however, do not produce, in general, the expected good results. Note also that in many practical applications the source p.d.f. is not known a priori and, thus, other approaches such as INFOMAX or ML can fail to separate the sources.

In this paper we will propose to explore another possibility. Since the marginal entropies of the outputs are difficult to estimate we will replace the measure of non-Gaussianity by a new one, which is simpler to evaluate, but still preserves the form of (4). We will replace each differential entropy by the respective $(1+\beta)$-order cumulant of the outputs,

$$h(y_i) \rightarrow \frac{|C_{y_i}^{1+\beta}|}{1+\beta}, \tag{5}$$

yielding the modified function

$$\Psi(\mathbf{G}) = \sum_{i=1}^{N} \frac{|C_{y_i}^{1+\beta}|}{1+\beta} - \log|\det(\mathbf{G})| - h(\mathbf{s}). \tag{6}$$

Other measures of non-Gaussianity could have been used, however, for the sake of simplicity, we have chosen to use non-normalized cumulants (see Appendix G for a brief description of the non-normalized cumulants in terms of moments).

It is interesting to observe that, although $\Psi(\mathbf{G})$ is no longer connected with the mutual information, when the mixing matrix is constrained to be orthogonal this function is the contrast for BSS proposed by Moreau and Macchi in [26]. But, in our approach, we prefer not to impose such a constraint because the exact orthogonality of the mixing matrix is not easy to obtain in the presence of additive and possibly, spatially correlated, Gaussian noise. The following theorem illustrates how, in comparison to the usual techniques for BSS, the separation solution is not a minimum or maximum of $\Psi(\mathbf{G})$ but a saddle point instead.

Theorem 1. *If all sources have $(1+\beta)$-order cumulants of unit modulo and β is an integer >1, source separation is always a saddle point of the function*

$$\Psi(\mathbf{G}) = \sum_{i=1}^{N} \frac{|C_{y_i}^{1+\beta}|}{1+\beta} - \log|\det(\mathbf{G})| - h(\mathbf{s}). \tag{7}$$

Proof. Let us first demonstrate that the gradient of $\Psi(\mathbf{G})$ vanishes at source separation (i.e., at $\mathbf{G} = \mathbf{I}$ or any permutation matrix). Using the cumulants properties we obtain in Appendix A that the gradient of $\Psi(\mathbf{G})$ with respect to $\mathbf{G}$ is

$$\frac{\partial \Psi(\mathbf{G})}{\partial \mathbf{G}} = \mathbf{S}_y^\beta \mathbf{C}_{y,s}^{\beta,1} - \mathbf{G}^{-\mathrm{T}}. \tag{8}$$

Equating the Gradient to zero, it is straightforward to see that our estimating equation is

$$\mathbf{S}_y^\beta \mathbf{C}_{y,y}^{\beta,1} = \mathbf{I}. \tag{9}$$

Since this equation only depends on the outputs, when there is no noise in the mixture, its solution will lead to an equivariant estimate of matrix $\mathbf{G}$ [11]. Moreover, when there is additive Gaussian noise in the mixture, the equivariant property will also hold in the asymptotic sense as the number of data points increases to infinity, thus, providing us with reliable estimates of the involved cross-cumulants.

At separation, $\mathbf{G} = \mathbf{I}$, and this equation will always hold true because $\mathbf{C}_{s,s}^{\beta,1} \mathbf{S}_s^\beta = \mathbf{I}$. It is interesting to point out the similarity between the estimating equation (9) and those that, coming from the maximum likelihood method, seek the diagonalization of a non-linear correlation matrix $E[\mathbf{f}(\mathbf{y})\mathbf{y}^\mathrm{T}] = \mathbf{I}$ where $\mathbf{f}(\cdot)$ is the score function [11,16]. The structure of both estimating equations is similar but, in our proposal, we employ cross-cumulant matrices instead of non-linear cross-correlations.

Next, let us examine the nature of the Hessian of $\Psi(\mathbf{G})$ at separation. The Hessian matrix, calculated in the Appendix A, is given by

$$\mathscr{H}\Psi(\mathbf{G}) = \frac{\partial}{\partial(vec\,\mathbf{G})^\mathrm{T}} \, vec\left(\frac{\partial\Psi(\mathbf{G})}{\partial(vec\,\mathbf{G})^\mathrm{T}}\right)^\mathrm{T}$$

$$= \beta\,diag(vec((\mathbf{G}^{\odot(\beta-1)})\mathbf{S}_s^\beta \mathbf{S}_y^\beta)) + \mathscr{K}_N((\mathbf{G}^{-1})^\mathrm{T} \otimes \mathbf{G}^{-1}), \tag{10}$$

where $\mathscr{K}_N$ is the commutation matrix, i.e., the permutation matrix that verifies $\mathscr{K}_N$ $vec(\mathbf{M}) = vec(\mathbf{M}^\mathrm{T})$.

At separation, we observe that

$$\mathscr{H}\Psi|_{\mathbf{G}=\mathbf{I}} = \beta\,diag(vec\,\mathbf{I}) + \mathscr{K}_N. \tag{11}$$

Since the eigenvalues of the Hessian matrix at separation (11) are $\{1+\beta, 1, -1\}$, the Hessian is neither positive definite or negative definite, proving that the separation is always a saddle point.

In order to graphically illustrate the previous result we show in Figs. 2 and 3 the shape of $\Psi(\mathbf{G})$ (for $\beta = 3$ and $N = 2$) in the neighborhood of the separation point. In this case, the Hessian matrix at separation takes the form

$$\mathscr{H}\Psi|_{\mathbf{G}=\mathbf{I}} = \begin{pmatrix} 1+\beta & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1+\beta \end{pmatrix}. \tag{12}$$
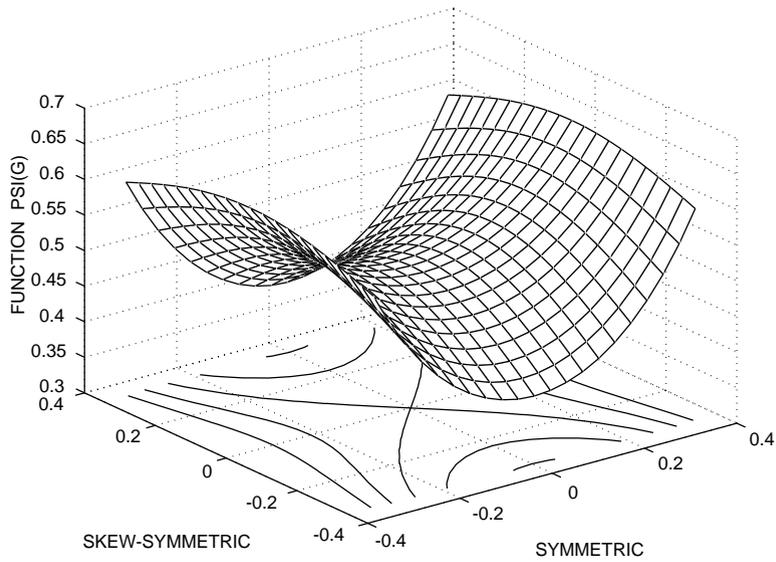
Fig. 2. Shape of the function $\Psi(\mathbf{G})$ obtained when $\beta = 3$ and the sources have the same kurtosis signs. The separation solution, which is located at the origin, is a saddle point.
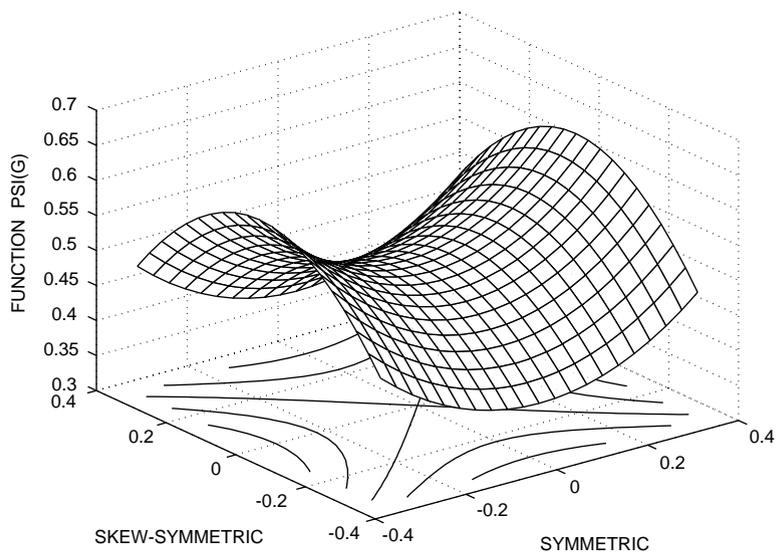

Fig. 3. Shape of the function $\Psi(\mathbf{G})$ obtained when $\beta = 3$ and the sources have different kurtosis signs. The separation solution is also a saddle point.

Since we are restricted to plotting the value of the function over a two dimensional subspace, we will use as input the subspace spanned by the following pair of variables (u, v). These are extracted, respectively, from a local symmetric and skew-symmetric parameterization of the 2×2 global transfer matrix $\mathbf{G}$. This parameterization, proposed in [10], is given by

$$\mathbf{G}(u, v) = \begin{pmatrix} \cosh(u) & \sinh(u) \\ \sinh(u) & \cosh(u) \end{pmatrix} \begin{pmatrix} \cos(v) & -\sin(v) \\ \sin(v) & \cos(v) \end{pmatrix}.$$

We can observe from both figures that the separation solution ($\mathbf{G}=\mathbf{I}$), which is attained at the origin of the graphs, is always a saddle point of $\Psi(\mathbf{G})$ regardless to the kurtosis signs of the sources.

3. Quasi-Newton algorithms

In this section we will present the first family of algorithms that converge to the separation solution. Since source separation is not a minimum or maximum of $\Psi(\mathbf{G})$ but a saddle point, we cannot use gradient-based approaches such as the conventional gradient or the natural gradient [2,11,13] to adapt the separating system. Instead, we propose to find the BSS solution by using a preconditioned method which employs the second-order information available at separation. In order to find the zeros of the gradient, we propose the utilization of a preconditioned iteration [21] of the form

$$vec\, \mathbf{G}^{(n+1)} = vec\, \mathbf{G}^{(n)} - \mu^{(n)}\, (\hat{\mathscr{H}}\Psi)^{-1}\, vec \left(\frac{\partial \Psi}{\partial \mathbf{G}} \right), \tag{13}$$

where $\hat{\mathscr{H}}\Psi$ is an approximation of the true Hessian matrix in the neighborhood of the separation. This class of numerical algorithm is a variant of the quasi-Newton *chord* methods [21].

Our proposal for the Hessian approximation is

$$\hat{\mathscr{H}}\Psi(\mathbf{G}) = \mathscr{K}_N((\mathbf{G}^{-1})^{\mathrm{T}} \otimes \mathbf{G}^{-1}), \tag{14}$$

which only differs from the true Hessian matrix in the diagonal terms. Moreover, this difference will become negligible at separation since the eigenvalues of the Hessian approximation at this point are those of the true Hessian but with unit modulo. On the other hand, as we will show in Appendices B and C, this difference is significant at the deceptive solutions of the estimating Eq. (9). For these, the sign of some eigenvalues of the Hessian approximation will change with respect to those of the true Hessian, avoiding, in this way, the possibility to converge to non-separating solutions.

Substituting (14) in the iteration (13) and returning to the matrix notation we arrive at the following algorithm:

$$\mathbf{G}^{(n+1)} = \mathbf{G}^{(n)} - \mu^{(n)}(\mathbf{C}_{y,y}^{1,\beta}\mathbf{S}_y^{\beta} - \mathbf{I})\mathbf{G}^{(n)}. \tag{15}$$

Multiplying (15) from the right by $\mathbf{A}^{-1}$ we obtain the iteration in terms of the separating system

$$\mathbf{B}^{(n+1)} = \mathbf{B}^{(n)} - \mu^{(n)}(\mathbf{C}_{y,y}^{1,\beta}\mathbf{S}_y^\beta - \mathbf{I})\mathbf{B}^{(n)}, \tag{16}$$

which we will denote as the cumulant-based iterative inversion (CII) algorithm. This name comes from the fact that this recursion can also be interpreted as a quasi-Newton algorithm that iteratively inverts a robust estimate of the mixing system. In our case, this estimate is given by $\hat{\mathbf{A}}(\mathbf{B}^{(n)}) = \mathbf{C}_{x,y}^{1,\beta}\mathbf{S}_y^\beta$ (see [15,16] for more details). Incidentally, it is interesting to observe that when we set $\beta = 1$ in the CII algorithm we find the following recursion:

$$\mathbf{B}^{(n+1)} = (\mathbf{I} - \mu^{(n)}(\mathbf{C}_{y,y}^{1,1} - \mathbf{I}))\mathbf{B}^{(n)} \tag{17}$$

that seeks the diagonalization of the symmetric correlation matrix $\mathbf{C}_{y,y}^{1,1}$. This is the globally stable decorrelation algorithm proposed by Almeida et al. in [1].

When implementing the algorithm CII it is useful to take into account that $\mathbf{S}_y^\beta = diag(sign(diag(\mathbf{C}_{y,y}^{1,\beta})))$ and that the cumulant matrix $\mathbf{C}_{y,y}^{1,\beta}$ can be obtained in terms of the moments of the outputs by using the relevant expressions from Appendix G. For instance, when we consider real signals and $\beta = 3$ the CII algorithm can be written in compact matrix form as

$$\mathbf{B}^{(n+1)} = \mathbf{B}^{(n)} - \mu^{(n)}.((E[\mathbf{y}(\mathbf{y}^{\odot 3})^\mathrm{T}] - 3\,E[\mathbf{y}\mathbf{y}^\mathrm{T}]\,diag(E[\mathbf{y}^{\odot 2}]))\mathbf{S}_y^3 - \mathbf{I})\mathbf{B}^{(n)}, \tag{18}$$

where the diagonal elements of the matrix $\mathbf{S}_y^3$ are defined as $[\mathbf{S}_y^3]_{ii} = \mathrm{sign}(E[y_i^4] - 3(E[y_i^2])^2)$.

Finally, it is possible to observe that the stochastic versions of the CII algorithms share a similar structural form with the Natural Gradient adaptation [2,11] except that now the positions of the linear and non-linear functions appear interchanged, i.e.,

$$\mathbf{B}^{(n+1)} = \mathbf{B}^{(n)} - \mu^{(n)}(\mathbf{y}\mathbf{g}(\mathbf{y})^\mathrm{T} - \mathbf{I})\mathbf{B}^{(n)}, \tag{19}$$

where $\mathbf{g}(\cdot)$ is the suitable function that acts componentwise on the output's vector. For instance, when $\beta = 3$, we find $[\mathbf{g}(\mathbf{y})]_i = (y_i^3 - 3y_i\hat{\sigma}_{y_i}^2)\hat{S}_{y_i}^3$ where $\hat{\sigma}_{y_i}^2$ and $\hat{S}_{y_i}^3$ are, respectively, the adaptive estimates of the power of the ith output and of the sign of its fourth-order cumulant.

3.1. Optimal choice of the step-size

Since the CII algorithm is of the quasi-Newton type, we should ensure that it always works in the regions where $\Psi(\mathbf{G})$ is continuous. This means that we should not reach or cross the discontinuities that occur when matrix $\mathbf{B}$ becomes singular. Since a necessary condition for $\mathbf{B}^{(n+1)} = (\mathbf{I} - \varDelta^{(n)})\mathbf{B}^{(n)}$ to be singular is that $\|\varDelta^{(n)}\| \geqslant 1$ for any chosen matrix norm, we only need to ensure that $\|\varDelta^{(n)}\| < 1$. Therefore, taking into account the triangular inequality

$$\|\mathbf{C}_{y,y}^{1,\beta}\mathbf{S}_y^\beta - \mathbf{I}\| \leqslant 1 + \|\mathbf{C}_{y,y}^{1,\beta}\mathbf{S}_y^\beta\| = 1 + \|\mathbf{C}_{y,y}^{1,\beta}\| \tag{20}$$

it is sufficient to choose

$$\mu^{(n)} = \min\left(\frac{2\eta}{1 + \eta\beta}, \frac{\eta}{1 + \eta\|\mathbf{C}_{y,y}^{1,\beta}\|} \right) \tag{21}$$

to avoid $\mathbf{B}^{(n+1)}$ becoming singular. Here $\eta < 1$ and the term $2\eta/(1+\eta\beta)$ will be justified later from the convergence properties of the algorithm.

3.2. Using other higher order information

One of the disadvantages of the CII algorithm is that it cannot be used when the $C_s^{1+\beta}$ cumulant is zero for any of the sources. To overcome this limitation we can use a set of indexes $\Omega = \{\beta_1, \ldots, \beta_{N_\Omega} : \beta_i \in \mathbb{N}^+, \beta_i \neq 1\}$ such that the following sum of cumulants $\sum_{\beta \in \Omega} |C_{s_i}^{1+\beta}|$ do not vanish for any of the sources. The existence of at least one possible set Ω is ensured by the non-Gaussianity of the sources. Then, instead of using a single cumulant, we can measure the non-Gaussianity in (4) by using a weighted sum of several cumulants of the outputs whose index β belong to the set Ω, i.e.,

$$h(y_i) \to \sum_{\beta \in \Omega} w_\beta \frac{|C_{y_i}^{1+\beta}|}{1 + \beta}, \tag{22}$$

where the positive weighting terms w_β are chosen so that $\sum_{\beta \in \Omega} w_\beta = 1$. With $1 \notin \Omega$ we will assume that second-order information is excluded from this weighted sum.

Following the same steps as before, it is straightforward to see that separation is still a saddle point of the resulting function. Similarly, we can derive the generalized and cumulant-based iterative inversion (GCII) algorithm to find this saddle point,

$$\mathbf{B}^{(n+1)} = \mathbf{B}^{(n)} - \mu^{(n)} \left(\sum_{\beta \in \Omega} w_\beta \mathbf{C}_{y,y}^{1,\beta} \mathbf{S}_y^\beta - \mathbf{I} \right) \mathbf{B}^{(n)}, \tag{23}$$

where

$$\mu^{(n)} = \min\left\{ \frac{2\eta}{1 + \eta \sum_{\beta \in \Omega} w_\beta \beta}, \frac{\eta}{1 + \eta\| \sum_{\beta \in \Omega} w_\beta \mathbf{C}_{y,y}^{1,\beta} \mathbf{S}_y^\beta \|} \right\} \tag{24}$$

and with $\eta < 1$ and $\sum_{\beta \in \Omega} w_\beta = 1$.

Since the extended algorithm exploits several cumulants matrices it will be more robust in the sense of reducing the probability that some bad choice of cumulant order results in near zero values of the weighted sum of cumulants for some sources. In addition, the use of several cumulants matrices increases the statistical information exploited by the algorithm.

It is outside of the scope of this paper to discuss the optimum weighting factors that should be used. Nevertheless, as a rule of thumb one can choose them to be proportionally inverse to the variance of the cumulant estimates.

The cost function (6) and the accompanying algorithms can be extended to applications for colored and non-stationary source signals. This can be done by replacing

the cost function (5) and the self-cumulants $C_{y_i}^{1+\beta}$ with cross-cumulants of the form $Cum(y_i[n], y_i[n - p_1], \ldots, y_i[n - p_\beta])$ where p_k, $k = 1, \ldots, \beta$, are suitably chosen time delays that satisfy $\sum_{\beta \in \Omega} Cum(s_i[n], s_i[n - p_1], \ldots, s_i[n - p_\beta]) \neq 0 \ \forall i$. In this case, it is straightforward to observe that the algorithms have the same structural form as before except for the fact that the entries of the cross-cumulant matrices should now be defined as follows: $C_{y_i,y_j}^{1,\beta} = Cum(y_i[n], y_j[n - p_1], \ldots, y_j[n - p_\beta])$.

4. Convergence properties

In this section we present two theorems about the convergence properties of the CII and GCII algorithms. The demonstration is given for the CII algorithm although the extension to the GCII version is straightforward.

Theorem 2. *Under the same conditions of the theorem I, the local convergence of CII and GCII algorithms, with a constant step-size, is almost isotropic and independent of the sources distribution.*

Proof. Let us start by rewriting the CII1 recursion (15) in vector form

$$vec\, \mathbf{G}^{(n+1)} = vec\, \mathbf{G}^{(n)} - \mu\, vec((\mathbf{C}_{y,y}^{1,\beta}\mathbf{S}_y^\beta - \mathbf{I})\mathbf{G}^{(n)}). \tag{25}$$

It is well known that the linearized version of any algorithm around the separation point determines its local convergence. The truncated first-order Taylor expansion of the CII iteration (25) at separation ($\mathbf{G} = \mathbf{I}$) is

$$vec(\mathbf{G}^{(n+1)}) = vec\, \mathbf{I} + \mathscr{J}\, vec(\mathbf{G}^{(n)} - \mathbf{I}), \tag{26}$$

where $\mathscr{J}$ is the overall Jacobian of the adaptation at separation. Defining $\mathscr{I}$ as the $N^2 \times N^2$ identity matrix, we can express this Jacobian matrix (see Appendix D) as

$$\mathscr{J} = \mathscr{I} - \mu(\mathscr{I} + \beta\, diag(vec\, \mathbf{I})). \tag{27}$$

We can observe from Eqs. (26) and (27) that the diagonal structure of the Jacobian matrix locally decouples the iteration for each element of $\mathbf{G}$. In fact, defining $\lambda_d = 1 - \mu(1 + \beta)$ and $\lambda_o = 1 - \mu$, the iteration (26) can be decomposed separately for the diagonal and non-diagonal elements of $\mathbf{G}$ as follows

$$(\mathbf{G}_{ii}^{(n+1)} - 1) = \lambda_d(\mathbf{G}_{ii}^{(n)} - 1),$$
$$\mathbf{G}_{ij}^{(n+1)} = \lambda_o \mathbf{G}_{ij}^{(n)}, \tag{28}$$

for all $i,j|_{i \neq j} = 1, \ldots, N$. We can see from Eq. (28) that the local convergence is determined by the $|\lambda_d|$ and $|\lambda_o|$ factors. Then, as long as the $(1 + \beta)$-order cumulants of the sources do not vanish, the local convergence does not depend on the sources p.d.f. In addition, the necessary and sufficient conditions for the asymptotic stability of

the algorithm are given by $|\lambda_o| < 1$ and $|\lambda_d| < 1$. Both conditions can be rewritten for the adaptation step-size as

$$0 < \mu < \min\left\{\frac{2}{1+\beta},\ 2\right\}. \tag{29}$$

Next, let us show that the CII algorithm exhibits almost isotropic convergence. The convergence of an algorithm is isotropic when the convergence rate is the same for all its adaptive elements. In our case, the convergence rate of the diagonal and off-diagonal terms of **G** is governed by the factors $|\lambda_d|$ and $|\lambda_o|$, respectively. From this result, and considering separately the behavior of the sets of the diagonal and non-diagonal elements of **G**, the elements of each of these sets converge isotropically (at the same rate) to the separation.

It is easy to see from Eq. (28) that the convergence rate will increase when $|\lambda_d|$ and $|\lambda_o|$ approach to zero. But since it is not possible to drive both factors to zero at the same time, we should find a compromise between the convergence speed of the diagonal and non-diagonal terms. Since near the solution the non-diagonal terms have greater importance for the separation, it is preferable to set $|\lambda_o| = 0$ which means that $\mu = 1$. However, we are limited by the local stability condition associated with the convergence of the diagonal terms, i.e., $\mu < 2/(1+\beta)$. Then, in order to choose the step-size closest to unity, while at the same time ensuring the convergence of the algorithm, we propose to set

$$\mu = \frac{2\eta}{1+\eta\beta} \tag{30}$$

with $\eta < 1$. When $\eta \approx 1$ (although it is always < 1) the convergence of the diagonal terms is under-damped and exhibits oscillations. Nevertheless, the convergence speed of the other terms increases.

We have shown that, by analyzing separately the diagonal and non-diagonal terms of **G**, and under the conditions of Theorem I, the local convergence of the algorithms is almost isotropic and independent of the source statistics. These properties are quite unusual with respect to the existing blind separation algorithms whose convergence and convergence rates usually depend on the source statistics. Moreover, as opposed to our approach, most of the competing algorithms [12,18,31] need the mixing matrix to be orthogonal. This is a condition that is difficult to impose when there is additive noise in the observations because it is usually sensitive and, thus, non-robust.

Theorem 3. *The convergence of CII and GCII algorithms will not be biased by the presence of Gaussian noise in the mixture.*

Proof. This theorem is straightforward to prove if we take into account that the CII and GCII algorithms use higher order cumulants as non-linearities and that the higher order cumulants of Gaussian stochastic processes are all zero.

5. Using second-order statistics

In general, since they contain useful information for the separation, it is desirable to incorporate second-order statistics into the adaptation rules, even though, when the sources are stationary they cannot carry out the blind separation by themselves. The main advantage of using second-order information is that their estimates have a lower variance than the estimates based on higher order statistics. But, unfortunately, using this information the algorithm will become biased when noise is present in the mixture.

For reasons of simplicity, let us consider in this section that the sources have constant unity variance instead of a certain constant cumulant modulo. Taking into account this consideration, the orthogonal constraint of the global matrix $\mathbf{G}$ will be tantamount to the decorrelation of the outputs, i.e., $\mathbf{C}_{y,y}^{1,1} = \mathbf{G}\mathbf{G}^{\mathrm{T}} = \mathbf{I}$.

In this section, to improve the separation in the noiseless case, we will present two methods to combine second-order and higher order information. The first method consists in the direct inclusion of the second-order cumulant into the weighted sum of cumulants. The second embeds the decorrelation adaptation (17) and the GCII algorithm into a single recursion.

The direct inclusion of the second-order cumulant into the weighted sum will be referred as the GCII+ algorithm. It can be obtained by substituting the set Ω in the cost function (22) and also in the recursion (23) by using the modified set $\Omega^+ = \{1, \beta_2, \ldots, \beta_{N_\Omega}: \beta_i \in \mathbb{N}^+\}$,

$$\mathbf{B}^{(n+1)} = \left(\mathbf{I} - \mu^{(n)} \left(\sum_{\beta \in \Omega^+} w_\beta \mathbf{C}_{y,y}^{1,\beta} \mathbf{S}_y^\beta - \mathbf{I}\right)\right) \mathbf{B}^{(n)}. \tag{31}$$

This modification keeps the separation solution as a saddle point of the cost function but changes the convergence behavior of the recursion. As shown in Appendix E, we loose the isotropic convergence property for non-i.i.d. sources. Furthermore, the adaptation step size has a slightly different constraint given by

$$\mu^{(n)} = \min\left(\frac{2\eta}{\beta_{\max}}, \frac{\eta}{1 + \eta\|\sum_{\beta \in \Omega^+} w_\beta \mathbf{C}_{y,y}^{1,\beta} \mathbf{S}_y^\beta\|}\right), \tag{32}$$

where $\eta < 1$ and $\beta_{\max} = \max_{\beta \in \Omega^+}\{\beta\}$.

A different way of incorporating second-order statistics uses serializing the decorrelation algorithm (17) with the GCII algorithm. This way of combining both adaptations is similar to the technique used by Cardoso and Laheld in [11]. Following this approach, we will split the separation system in two parts $\mathbf{B} = \mathbf{U}\mathbf{W}$. The first system, with outputs $\mathbf{z} = \mathbf{W}\mathbf{x}$, will account for the prewhitening of the observations. Thus, the decorrelation adaptation

$$\mathbf{W}^{(n+1)} = (\mathbf{I} - \mu_1^{(n)}(\mathbf{C}_{z,z}^{1,1} - \mathbf{I}))\mathbf{W}^{(n)} \tag{33}$$

will be suited for this objective. On the other hand, the second system $\mathbf{U}$, whose outputs are $\mathbf{y} = \mathbf{U}\mathbf{z}$, will provide the adequate rotation that separates the sources. This

can be accomplished, for instance (when $\beta > 1$), by maximizing the contrast function proposed by Moreau and Macchi in [26]

$$\psi_\beta = \sum_{i=1}^{N} \frac{|C_{y_i}^{1+\beta}|}{1+\beta} \quad \text{subject } to \ \mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}, \tag{34}$$

or, in general, maximizing

$$\Psi = \sum_{\beta \in \Omega} w_\beta \psi_\beta \quad \text{subject } to \ \mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}. \tag{35}$$

The maximization of Ψ can be carried out in the Stiefel manifold of orthogonal matrices using the natural gradient ascent direction [3]

$$\tilde{\nabla}_{\mathbf{U}} \Psi = \frac{\partial \Psi(\mathbf{U})}{\partial \mathbf{U}} - \mathbf{U} \left(\frac{\partial \Psi(\mathbf{U})}{\partial \mathbf{U}} \right)^{\mathrm{T}} \mathbf{U}$$

$$= \sum_{\beta \in \Omega} w_\beta (\mathbf{S}_y \mathbf{C}_{y,y}^{3,1} - \mathbf{C}_{y,y}^{1,3} \mathbf{S}_y) \mathbf{U}^{(n)}, \tag{36}$$

which results in the following algorithm:

$$\mathbf{U}^{(n+1)} = \mathbf{U}^{(n)} + \mu_2^{(n)} \sum_{\beta \in \Omega} w_\beta (\mathbf{S}_y \mathbf{C}_{y,y}^{3,1} - \mathbf{C}_{y,y}^{1,3} \mathbf{S}_y) \mathbf{U}^{(n)}. \tag{37}$$

Since $\mathbf{B}^{(n)} = \mathbf{U}^{(n)} \mathbf{W}^{(n)}$ we can serialize both algorithms [11]

$$\mathbf{B}^{(n+1)} = \mathbf{U}^{(n+1)} \mathbf{W}^{(n+1)} \tag{38}$$

$$= \left(\mathbf{I} + \mu_2^{(n)} \sum_{\beta \in \Omega} w_\beta (\mathbf{S}_y \mathbf{C}_{y,y}^{3,1} - \mathbf{C}_{y,y}^{1,3} \mathbf{S}_y) \right) \mathbf{U}^{(n)}.$$

$$(\mathbf{I} - \mu_1^{(n)} (\mathbf{C}_{z,z}^{1,1} - \mathbf{I})) \mathbf{W}^{(n)} \tag{39}$$

$$= \left(\mathbf{I} - \mu_1^{(n)} (\mathbf{C}_{y,y}^{1,1} - \mathbf{I}) + \mu_2^{(n)} \sum_{\beta \in \Omega} w_\beta (\mathbf{S}_y \mathbf{C}_{y,y}^{3,1} - \mathbf{C}_{y,y}^{1,3} \mathbf{S}_y) \right) \mathbf{B}^{(n)}$$

$$+ O(\mu_1^{(n)} \mu_2^{(n)}) \tag{40}$$

and, neglecting the term $O(\mu_1^{(n)} \mu_2^{(n)})$, we obtain a new cumulant-based EASI algorithm

$$\mathbf{B}^{(n+1)} = \left(\mathbf{I} - \mu_1^{(n)} (\mathbf{C}_{y,y}^{1,1} - \mathbf{I}) + \mu_2^{(n)} \sum_{\beta \in \Omega} w_\beta (\mathbf{S}_y \mathbf{C}_{y,y}^{3,1} - \mathbf{C}_{y,y}^{1,3} \mathbf{S}_y) \right) \mathbf{B}^{(n)}, \tag{41}$$

which we will denote as CEASI.

So that the step sizes, $\mu_1^{(n)}$ and $\mu_2^{(n)}$, prevent the separating system from reaching singularities; it is sufficient that they do not exceed the following bound

$$\mu_0^{(n)} = \frac{\eta}{1 + \eta \|\mathbf{C}_{y,y}^{1,1} - \sum_{\beta \in \Omega} w_\beta (\mathbf{S}_y \mathbf{C}_{y,y}^{3,1} - \mathbf{C}_{y,y}^{1,3} \mathbf{S}_y)\|} \tag{42}$$

for any given matrix norm and $\eta < 1$.

Since it imposes a fixed variance for the outputs, one of the disadvantages of this way of incorporating second-order statistics is that we lose the isotropic convergence property for non-i.i.d. sources. Then, the convergence could slow down when there are large differences in the weighted sum of cumulants between the sources. In spite of this, if the adaptation steps-sizes are properly chosen, the method will always be locally convergent to the separation. This issue is addressed in Appendix F where we calculate the local stability conditions for the algorithm and the resulting constraints for the step-size. Simple expressions for the step-size that satisfy these constraints and at the same time are close to their locally optimal values are given by

$$\mu_{1*}^{(n)} = \frac{1}{2}, \tag{43}$$

$$\mu_{2*}^{(n)} = \frac{1}{2} \left(\max_{i=1,\dots,N} \left\{ \sum_{\beta \in \Omega} w_\beta |C_{y_i}^{1+\beta}| \right\} \right)^{-1}. \tag{44}$$

Then, combining all the constraints and recommendations for the step-sizes, we finally choose

$$\mu_1^{(n)} = \min\{\mu_0^{(n)}, \ \mu_{1*}\},$$

$$\mu_2^{(n)} = \min\{\mu_0^{(n)}, \ \mu_{2*}^{(n)}\}. \tag{45}$$

6. Extensions

6.1. Complex sources and mixtures

The extension of the previous algorithms to the case of complex sources and mixtures is obtained by simply replacing the transpose operator $(\cdot)^\mathrm{T}$ with the Hermitian or conjugate transpose operator $(\cdot)^\mathrm{H}$ and by changing the definition of the cumulants and cross-cumulants. Whenever $1 + \beta$ is an even number, it is possible to define the cumulants as

$$C_{y_i}^{1+\beta} = Cum(\underbrace{y_i, \dots, y_i}_{\frac{1+\beta}{2}}, \underbrace{y_i^*, \dots, y_i^*}_{\frac{1+\beta}{2}}). \tag{46}$$

It is easy to show that they are always real. This makes possible to replace the real gradient operator used in the derivations by the complex gradient operator defined in [7], with exception of the interchange of the Hermitian and transpose operators

mentioned above. These modifications yield to the same results. We should also note that the cross-cumulant $C_{y_i,y_j}^{1,\beta}$ is now defined as

$$C_{y_i,y_j}^{1,\beta} = Cum(y_i, \underbrace{y_j, \ldots, y_j}_{\frac{1+\beta}{2}-1}, \underbrace{y_j^*, y_j^*, \ldots, y_j^*}_{\frac{1+\beta}{2}}). \tag{47}$$

6.2. More sensors than sources

The derivations of the proposed algorithms and its stability analysis has been performed in terms of the global transfer matrix $\mathbf{G}$. As long as this matrix remains square and non-singular, these derivations in terms of $\mathbf{G}$ still fully apply. This also includes the case where the number of sources and outputs (recovered sources) coincide with N and the number of sensors M is greater than the number of sources ($M > N$).

As an example, let us take the CII adaptation derived in terms of $\mathbf{G}$ as

$$\mathbf{G}^{(n+1)} = (\mathbf{I} - \mu^{(n)}(\mathbf{C}_{y,y}^{1,\beta}\mathbf{S}_y^\beta - \mathbf{I}))\mathbf{G}^{(n)}. \tag{48}$$

We can post-multiply it by the pseudo-inverse of the mixing matrix $\mathbf{A}^+ = (\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H$ to obtain

$$\mathbf{B}^{(n+1)}\mathbf{P_A} = (\mathbf{I} - \mu^{(n)}(\mathbf{C}_{y,y}^{1,\beta}\mathbf{S}_y^\beta - \mathbf{I}))\mathbf{B}^{(n)}\mathbf{P_A}, \tag{49}$$

where $\mathbf{P_A} = \mathbf{A}\mathbf{A}^+$ is the projection matrix onto the subspace spanned by the columns of $\mathbf{A}$. Thus, the algorithm is now defined in terms of the separation matrix $\mathbf{B}^{(n+1)}\mathbf{P_A}$ instead of $\mathbf{B}^{(n+1)}$. Nevertheless, we can remove the projection term $\mathbf{P_A}$ to obtain the CII2 algorithm for the non-square matrix case:

$$\mathbf{B}^{(n+1)} = (\mathbf{I} - \mu^{(n)}(\mathbf{C}_{y,y}^{1,\beta}\mathbf{S}_y^\beta - \mathbf{I}))\mathbf{B}^{(n)}. \tag{50}$$

This removal does not affect the signal component of the outputs since $\mathbf{P_A}\mathbf{A} = \mathbf{I}$ and, therefore, the separation is still performed. Similarly, we can also extend this reasoning for the GCII, GCII+ and CEASI algorithms in the case of more sensors than sources.

6.3. MMSE separating system

In the previous sections we have analyzed the ability of the CII and GCII algorithms to converge towards the point $\mathbf{B} = \mathbf{A}^{-1}$ and we have seen that convergence at this point is not biased by the presence of Gaussian noise. In the absence of noise, the separating system $\mathbf{B} = \mathbf{A}^{-1}$ is optimum because the sources are perfectly recovered at the output. However, when there is noise, this separating system may yield to an amplification of the noise at the outputs, especially if the mixing matrix is ill-conditioned. This phenomenon is also observed in channel equalization experiments when the zero-forcing criterion is used [25].

To avoid this limitation we can use the minimum mean square error (MMSE) separating system that exhibits a desirable balance between source separation and noise enhancement. The MMSE separating system, obtained from the classic orthogonality

principle when $(\mathbf{y} - \mathbf{s})$ and $\mathbf{x}$ are orthogonal, is given by

$$\mathbf{B}_{\text{MMSE}} = \mathbf{C}_{s,x}^{1,1}(\mathbf{C}_{x,x}^{1,1})^{-1}, \tag{51}$$

where $\mathbf{C}_{s,x}^{1,1}$ is the cross-correlation matrix between the sources and the observations, while $\mathbf{C}_{x,x}^{1,1}$ is the autocorrelation matrix of the observations.

An issue that should be taken into account is that the proposed algorithms will be able to separate the source signals up to a scaling factor (associated to a diagonal matrix Λ) and a possible permutation of the sources $\mathbf{P}$ in such a way that $\mathbf{G} = \Lambda \mathbf{P}$. In order to completely eliminate the indeterminacies of the separation matrix we can assume that by using some indirect knowledge (such as the shape of the p.d.f. or the value of the cumulants of order $1 + \beta$ of the sources) we can determine this permutation matrix and the mentioned scaling factor. Otherwise, these indeterminacies will propagate also to the MMSE matrix.

Since the noise and the sources are mutually independent, the MMSE solution can be rewritten in terms of the sources correlation matrix $\mathbf{C}_{s,s}^{1,1}$ as

$$\mathbf{B}_{\text{MMSE}} = \mathbf{C}_{s,s}^{1,1}\mathbf{P}^{\text{T}}\Lambda^{*}\mathbf{B}^{-\text{H}}(\mathbf{C}_{x,x}^{1,1})^{-1} \tag{52}$$

or, using the matrix inversion lemma, in terms of the noise correlation matrix $\mathbf{C}_{e,e}^{1,1}$ as

$$\mathbf{B}_{\text{MMSE}} = \mathbf{P}^{\text{T}}\Lambda^{-1}\mathbf{B}(\mathbf{I} - \mathbf{C}_{e,e}^{1,1}(\mathbf{C}_{x,x}^{1,1})^{-1}). \tag{53}$$

7. Simulations

Computer simulations were carried out to illustrate the performance of the proposed algorithms. In order to better represent the simulation results, we chose to use complex signals. This way we can observe the degree of separation that has been reached from the plot of their values on the complex plane.

In a first computer experiment we chose three sources with different p.d.f.s. The first signal is a 16-QAM (a communications signal used in Quadrature Amplitude Modulation digital transmissions [25] whose constellation has 16 complex discrete values or symbols) and has negative kurtosis (fourth-order cumulant). The second one is a 4-QAM signal whose kurtosis is also negative and the third signal is an asymmetric, real and zero mean source with positive kurtosis. This last source was generated by raising the samples obtained from a uniform random variable $U[0,1]$ to the seventh power and then subtracting its mean $(1/8)$. All the sources are normalized to a unit power. We define the mixing matrix in terms of the complex number $c = 1 + j$ as

$$\mathbf{A} = \begin{pmatrix} 1 & 0.8c & 0.5c^{*} \\ 0.5c^{*} & 1 & 0.7c \\ 0.8c^{*} & 0.5c & 1 \end{pmatrix}. \tag{54}$$
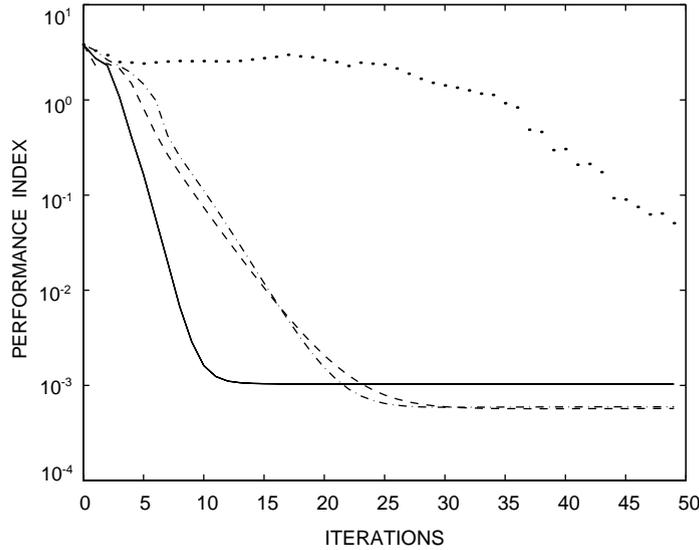
Fig. 4. Performance index versus iterations in the noiseless case: CII continuous line, GCII+ dashed line, CEASI dashed–dotted line, Extended-Infomax dotted line.

The algorithm performance is measured in terms of the following index P_{index}:

$$P_{\text{index}} = \sum_{i=1}^{N} \left(\sum_{j=1}^{N} \frac{|G_{ij}|^2}{\max_l \{|G_{il}|^2\}} - 1 \right) + \sum_{j=1}^{N} \left(\sum_{i=1}^{N} \frac{|G_{ij}|^2}{\max_l \{|G_{lj}|^2\}} - 1 \right). \qquad (55)$$

The three algorithms CII, GCII+ and CEASI, have been implemented in a batch way in order to obtain an accurate comparison between them and also to verify their theoretical properties. Nevertheless, on-line adaptive implementations for them are also feasible. The Extended-Infomax algorithm [17,24] has also been implemented for comparison purposes. The cross-cumulants of the outputs have been estimated in terms of the moments as shown in Appendix G and these statistics were calculated from a data block of 5000 observations. The common parameters are $\eta = 0.9$, $\Omega = \{3\}$ for the CII and CEASI algorithms, $\Omega^+ = \{1, 3\}$ for the GCII+ algorithm, and the 1-norm used for the calculation of the step-sizes.

The convergence results are shown in Fig. 4 for the noiseless case. We can observe that even though the three sources have very different p.d.f. (with different kurtosis sign) the algorithms all converge to the separation. However, the CII algorithm is the one that presents greater convergence speed, reaching the separation in about 10 iterations. This greater speed is a consequence of the local isotropic convergence of this algorithm. The fact that asymptotic performance reached by the CII algorithm is slightly poorer than that of the others is because this algorithm only exploits the information present in the statistic $\mathbf{C}_{y,y}^{1,3}$, whereas the algorithms GCII+ and CEASI
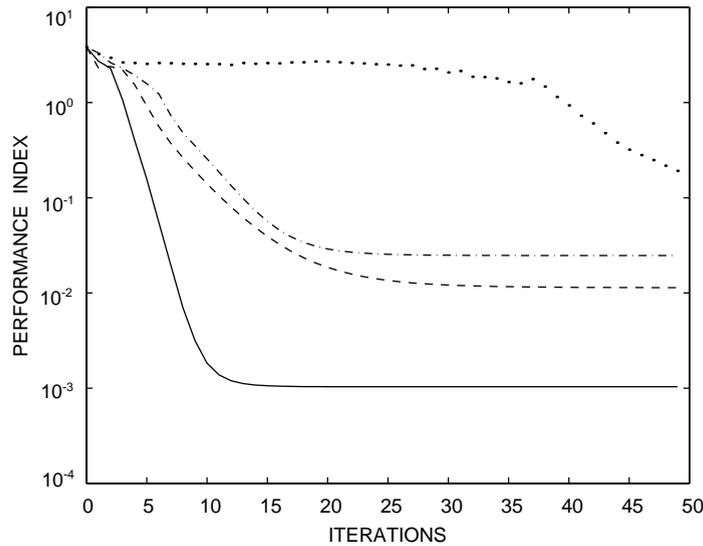
Fig. 5. Performance index versus iterations in the noisy case: CII continuous line, GCII+ dashed line, CEASI dashed–dotted line, Extended-Infomax dotted line.

Table 1
Minimum mean square error between the sources and the different signals

Signals	$\mathbf{x}$	$\mathbf{y}$	$\mathbf{y}_{MMSE}$	$\mathbf{y}_{sc}$
MMSE	1.35	0.45	0.38	0.02

also use second-order statistics. We additionally observe that the Extended-Infomax algorithm converges much slower than the proposed cumulant based algorithms.

The results corresponding to the noisy case are presented in Fig. 5. The input SNR is only 5 dB. The convergence of the CII algorithm is equal to the previous case, thus corroborating its theoretical unbiasedness with respect to the Gaussian noise. This does not occur for the other algorithms whose performance degrades in the presence of noise. Fig. 6 shows the results obtained for the CII algorithm in the previous experiment. The first column depicts the sources, the second column the observations, the third column the outputs after convergence and the fourth column the signal component of each output. It can be seen that, even though the outputs are distorted due to the noise, the algorithm has been able to separate the sources.

If we assume we know the sources or the noise power, we can find the separation matrix associated to the MMSE criterion. Let us assume that we are able to remove the BSS indeterminacies, i.e., we know Λ and $\mathbf{P}$. Table 1 indicates the resultant MMSE of the following signals: the observations ($\mathbf{x}$), the outputs after separation ($\mathbf{y}$), the outputs obtained after applying the MMSE criterion ($\mathbf{y}_{MMSE}$), and the signal component of the outputs after separation ($\mathbf{y}_{sc}$). This result, combined with several other simulations,
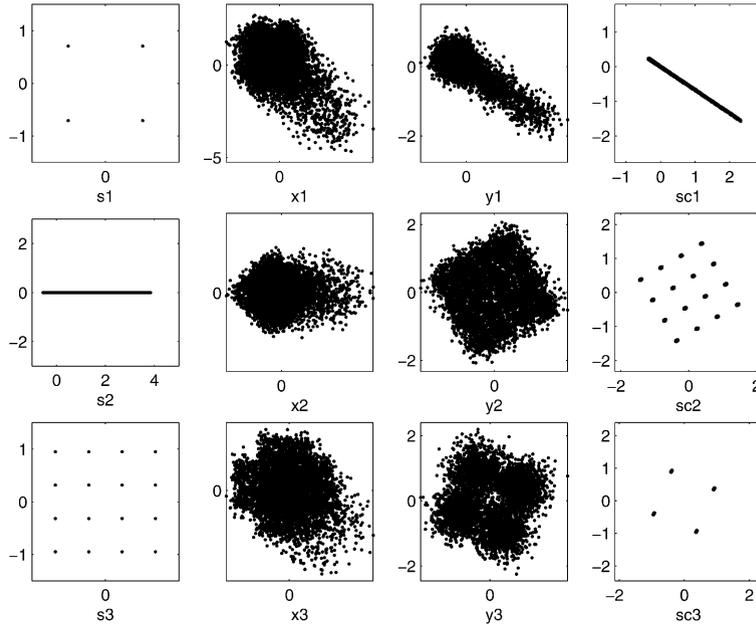
Fig. 6. First column: source signals. Second column: observations. Third column: outputs after convergence. Fourth column: signal component in the outputs.

confirms the hypothesis that for moderate/high signal-to-noise ratios and well-conditioned mixing matrices the MMSE solution is usually close to the separation solution. This fact can be also inferred from Eq. (53).

Finally, in order to corroborate the convergence analysis of the algorithms, we simulated the response of the algorithms in a situation where the number of sensors is greater than the number of sources and the sources are non-identically distributed. We consider eight sensors and a set of five non-i.i.d. sources with unit variance and whose kurtosis are chosen randomly during each simulation from an uniform distribution within the interval $[-2, 2]$. We also consider no noise in the model as well as mixing and separating matrices whose elements are generated randomly during each simulation. The parameters chosen for the algorithms were the same as before. We will assume in the simulations that infinite length data is available so that the estimates of the involved cross-cumulants have zero variance. The results of 50 different simulations for each algorithm are shown in Figs. 7–9 . We can see from the figures that all the algorithms are convergent to the separating point. Fig. 7 shows that for the CII algorithms, the asymptotic rate of convergence does not depend on the mixing nor on the source distribution. This fact corroborates both the isotropic local convergence of the CII algorithm and its independence with respect to the sources statistics. As can be seen from Figs. 8 and 9, the GCII+ and CEASI algorithms lose the isotropic local convergence when the sources are non-i.i.d.
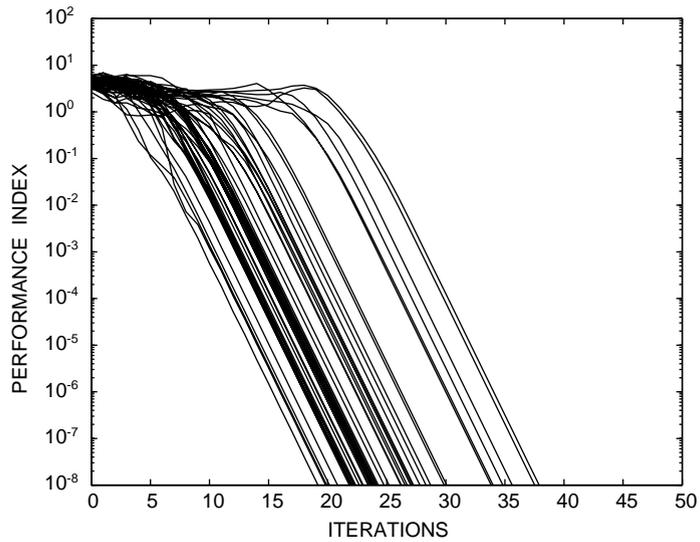
Fig. 7. Convergence of fifty random simulations of the CII algorithm for different sets of five sources and eight sensors, where each source has a random kurtosis within the interval $[-2, 2]$.
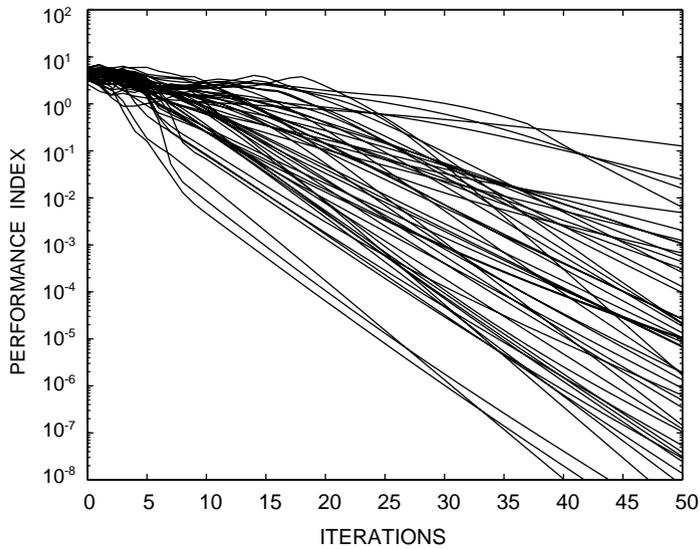

Fig. 8. Convergence of fifty random simulations of the GCII+ algorithm for different sets of five sources and eight sensors, where each source has a random kurtosis within the interval $[-2, 2]$.
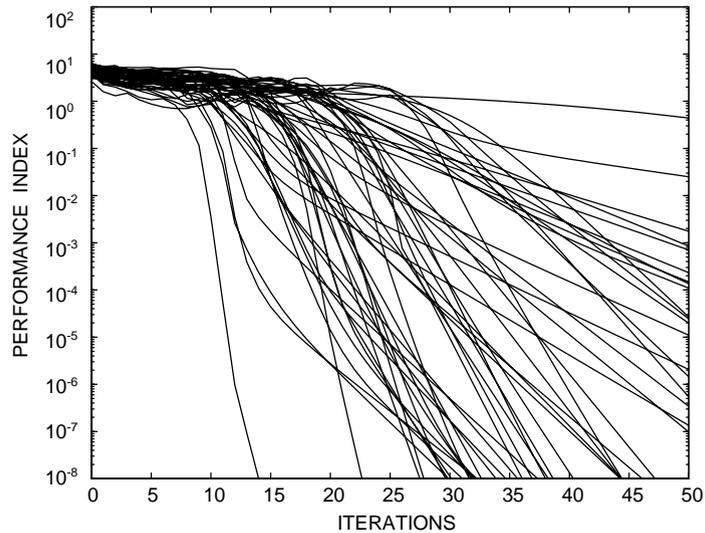
Fig. 9. Convergence of fifty random simulations of the CEASI algorithm for different sets of five sources and eight sensors, where each source has a random kurtosis within the interval $[-2, 2]$.

8. Conclusions

We have presented a new approach to blind source separation using cumulants. The method is motivated by the fact that, almost regardless of the distribution of the sources, separation can be achieved by seeking a saddle point of a cumulant-based cost function. This way we overcome the limitations of the INFOMAX and Maximum Likelihood approaches that may not work when the marginal p.d.f.s of the sources are unknown. To determine the saddle point of the proposed cost function we have used a quasi-Newton method that yields several families of algorithms for BSS. The convergence properties of these algorithms have been investigated and their local convergence has been demonstrated. The family of algorithms which only use higher order cumulants, as opposed to many of the existing algorithms, avoids the need of using sensitive preprocessing stages to orthogonalize the mixing matrix. This leads to more robust estimates. Moreover, the algorithms in this family also exhibit the following interesting properties: local isotropic convergence, a local convergence behavior almost independent of the source statistics and asymptotic equivariance of the estimates even in the presence of additive Gaussian noise.

9. For further reading

The following reference may also be of interest to the reader: [23].

Appendix A. Gradient and Hessian of $\Psi(\mathbf{G})$

Let us calculate the gradient of the function $\Psi(\mathbf{G})$ defined in (7). First note that it can be rewritten as

$$\Psi(\mathbf{G}) = \frac{1}{1+\beta}\, tr\{\mathbf{S}_y^\beta \mathbf{C}_{y,y}^{1,\beta}\} - \log|\det(\mathbf{G})| - h(\mathbf{s}). \tag{A.1}$$

Taking into account that

$$\mathbf{C}_{y,y}^{1,\beta} = \mathbf{G}\mathbf{C}_{s,s}^{1,\beta}(\mathbf{G}^{\odot\beta})^{\mathrm{T}} \tag{A.2}$$

and the multi-linear property of the cumulants [27], it is easy to compute the first differential of $\Psi(\mathbf{G})$ as

$$\mathrm{d}\Psi(\mathbf{G}) = tr\{(\mathbf{C}_{s,y}^{1,\beta}\mathbf{S}_y^\beta - \mathbf{G}^{-1})\,\mathrm{d}\mathbf{G}\}. \tag{A.3}$$

Therefore, the gradient of $\Psi(\mathbf{G})$ with respect to $\mathbf{G}$ is given by

$$\frac{\partial\Psi(\mathbf{G})}{\partial\mathbf{G}} = \mathbf{S}_y^\beta \mathbf{C}_{y,s}^{\beta,1} - \mathbf{G}^{-\mathrm{T}}. \tag{A.4}$$

In order to compute the Hessian we take the second differential,

$$\mathrm{d}^2\Psi(\mathbf{G}) = tr\{(\mathbf{C}_{s,s}^{1,\beta}d(\mathbf{G}^{\odot\beta})^{\mathrm{T}}\mathbf{S}_y^\beta - \mathrm{d}\mathbf{G}^{-1})\,\mathrm{d}\mathbf{G}\} \tag{A.5}$$

$$= tr\{(\beta\ (\mathbf{C}_{s,y}^{2,\beta-1}\odot\mathrm{d}\mathbf{G}^{\mathrm{T}})\mathbf{S}_y^\beta + \mathbf{G}^{-1}\mathrm{d}\mathbf{G}G^{-1})\,\mathrm{d}\mathbf{G}\} \tag{A.6}$$

$$= (vec\,\mathrm{d}\mathbf{G})^{\mathrm{T}}(\beta\,diag\,vec(\mathbf{G}^{\odot(\beta-1)}\mathbf{S}_s^\beta\mathbf{S}_y^\beta))\,vec\,\mathrm{d}\mathbf{G}$$
$$+ tr\{\mathbf{G}^{-1}\mathrm{d}\mathbf{G}G^{-1}\mathrm{d}\mathbf{G}\}. \tag{A.7}$$

Thus, after some straightforward manipulations, we obtain the Hessian matrix

$$\mathscr{H}\Psi(\mathbf{G}) = \beta\,diag(vec(\mathbf{G}^{\odot(\beta-1)}\mathbf{S}_s^\beta\mathbf{S}_y^\beta)) + \mathscr{K}_N((\mathbf{G}^{-1})^{\mathrm{T}}\otimes\mathbf{G}^{-1}), \tag{A.8}$$

where $\mathscr{K}_N$ is the permutation matrix for which $\mathscr{K}_N\,vec(\mathbf{M}) = vec(\mathbf{M}^{\mathrm{T}})$.

Appendix B. Solutions of the estimating Eq. (9)

The estimating equation we propose to solve is given by

$$\mathbf{S}_y^\beta\mathbf{C}_{y,y}^{\beta,1} = \mathbf{S}_y^\beta\mathbf{G}\mathbf{S}_s^\beta(\mathbf{G}^{\odot\beta})^{\mathrm{T}} = \mathbf{I}. \tag{B.1}$$

Let us denote $\mathbf{g}_{i:}$ as the ith row of matrix $\mathbf{G}$. Then, we can rewrite equation (B.1) as the following set of conditions for all $i,j|_{i\neq j} = 1,\ldots,N$:

$$\langle\mathbf{g}_{i:}^{\odot\beta},\mathbf{g}_{j:}\mathbf{S}_s^\beta\rangle = 0, \tag{B.2}$$

$$\langle\mathbf{S}_{s_i}^\beta\mathbf{g}_{i:}^{\odot\beta},\mathbf{g}_{i:}\mathbf{S}_s^\beta\rangle = 1, \tag{B.3}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of two vectors. From (B.1) we observe that $\mathbf{GS}_s^\beta$ has full rank, thus, its rows form a set of N linearly independent vectors that span the whole real space of dimension N. Taking this into account, the only possible solutions to the Eqs. (B.2) and (B.3) are of the form

$$\mathbf{g}_{i:}^{\odot\beta} = \alpha_i \mathbf{g}_{i:} \mathbf{S}_s^\beta \tag{B.4}$$

$$= (\|\mathbf{g}_{i:}\|^{-2} \mathbf{S}_{y_i}^\beta) \mathbf{g}_{i:} \mathbf{S}_s^\beta. \tag{B.5}$$

All the non-zero elements of the ith row of $\mathbf{G}$ must have modulo $(k_i)^{1/(\beta-1)}$ where k_i is the number of the elements in $\mathbf{g}_{i:}$ that are non-zero.

Let us only analyze, for the sake of brevity, the case of β odd. From Eq. (B.4) we observe that $g_{ij} = 0$ if $\mathbf{S}_{y_i} \neq \mathbf{S}_{s_j} \ \forall i, j$, thus, we do not need to worry about the signs of the cumulants because they will have no effect on the non-null elements of $\mathbf{G}$, and we can remove them in our study.

With this idea in mind, the candidates to solve (B.1) can be rewritten as

$$\mathbf{G} = diag([k_1, \ldots, k_n])^{-1/(\beta+1)} \mathbf{H}, \tag{B.6}$$

where all elements of $\mathbf{H}$ belong to the set $\{-1, 0, -1\}$. Substituting $\mathbf{G}$ into the estimating Eq. (B.1) yields the following equivalent condition:

$$\mathbf{HH}^T = diag([k_1, \ldots, k_n]). \tag{B.7}$$

The solutions of this equation are those matrices $\mathbf{H}$ which (up to row permutations) are diagonal by blocks, and whose m blocks are Hadamard matrices of respective dimensions $l_i \times l_i$ where $i = 1, \ldots, m$.

The Hadamard matrix of size one is just $\mathbf{H}_{(1)} = 1$. Given the Hadamard matrix of dimension $n \times n$ one can construct another matrix of dimension $2n \times 2n$ using the following recursive procedure:

$$\mathbf{H}_{(2n)} = \left(\begin{array}{c|c} \mathbf{H}_{(n)} & \mathbf{H}_{(n)} \\ \hline \mathbf{H}_{(n)} & -\mathbf{H}_{(n)} \end{array} \right). \tag{B.8}$$

A necessary condition for the existence of a Hadamard matrix of a given dimension $n \times n$ (where $n > 2$) is that n be divisible by 4.

Appendix C. Stability analysis of the deceptive solutions of (9)

When the solution of the estimating Eq. (B.7) involves a single Hadamard matrix $l_1 = N$, it is easy to obtain the minimum eigenvalue of the Hessian matrix (10), which is given by

$$\lambda_{\min} = (\beta - 1) \cdot |N^{-\frac{2}{\beta+1}}| > 0. \tag{C.1}$$

As a consequence of $\beta > 1$, we can observe that the true Hessian at this point will be positive definite regardless of N, indicating that this solution is a minimum of $\Psi(\mathbf{G})$
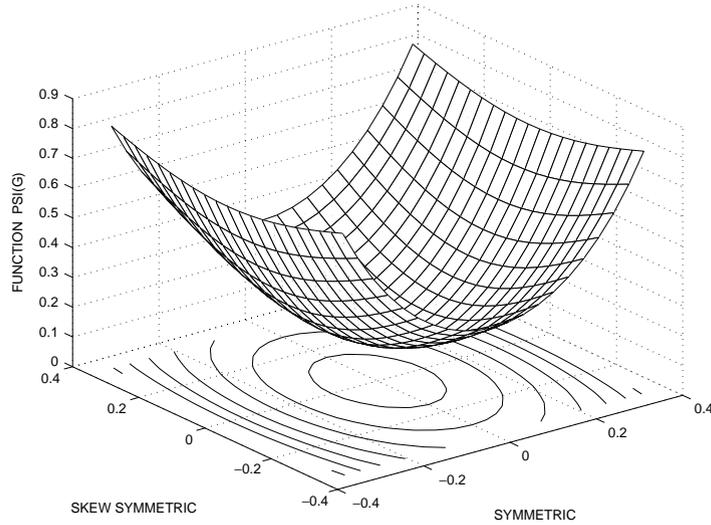
Fig. 10. Shape of the function $\Psi(\mathbf{G})$ in the vicinity of the deceptive solution $\mathbf{G} = (1/2)^{1/(1+\beta)}[1, 1; 1, -1]$ when $\beta = 3$. The deceptive solution, located after a local parametrization at the origin, is a minimum of $\Psi(\mathbf{G})$.

(see Fig. 10). Since the proposed algorithm was designed to converge to saddle points, this kind of deceptive critical points of the estimating equation cannot be stable points for the algorithm.

On the other hand, when the solution of (B.7) is not formed by a single Hadamard matrix, but by a combination of several Hadamard matrices, then $\mathbf{G}$ (up to row permutations) is also diagonal by blocks with the same structure of $\mathbf{H}$, i.e.,

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{G}_l \end{pmatrix}. \tag{C.2}$$

Note that at the separation solution $(l = N)$ all the diagonal blocks are of dimension 1.

If we analyze the behavior of $\Psi(\cdot)$ only with respect to those elements of each block $\mathbf{G}_i$ (while keeping the remaining coefficients constant) we observe from (C.1) that the critical points that satisfy the estimating equation are local minima of $\Psi(\mathbf{G})$ in their respective subspaces. This is easy to prove since the Hessian matrices of $\Psi(\mathbf{G})$ with respect to the elements of each of the blocks $\mathbf{G}_i$, $i = 1, \dots, l$, are positive definite at the critical points of the estimating equation. Thus, at any deceptive solution $(l \neq N)$ the true Hessian will differ with respect to the approximated Hessian in the sign of some eigenvalues and, therefore, it will be an unstable point for the proposed algorithm. Only for the special case of the separation, i.e., when all the diagonal blocks are of

size 1, the true Hessian and the modified Hessian share the same eigenvalues signs and, therefore, only at the separation is the solution stable.

Appendix D. Jacobian of the adaptation CII at the separation

Given the CII adaptation in terms of the global system

$$\mathbf{G}^{(n+1)} = \mathbf{G}^{(n)} - \mu(\mathbf{C}_{y,y}^{1,\beta}\mathbf{S}_y^\beta - \mathbf{I})\mathbf{G}^{(n)} \tag{D.1}$$

we can define the matrix ε as a small perturbation of the global transfer matrix from the separation, i.e., $\varepsilon = \mathbf{G}^{(n)} - \mathbf{I}$. Substituting $\mathbf{G}^{(n)}$ in terms of ε and, taking into account that close to the separation $\mathbf{S}_y^\beta = \mathbf{S}_s^\beta$, we can expand the iteration as

$$\mathbf{G}^{(n+1)} = (\mathbf{I} + \varepsilon) - \mu((\mathbf{I} + \varepsilon)\mathbf{C}_{s,s}^{1,\beta}((\mathbf{I} + \varepsilon)^{\odot\beta})^{\mathrm{T}}\mathbf{S}_y^\beta - \mathbf{I})(\mathbf{I} + \varepsilon) \tag{D.2}$$

$$= \mathbf{I} + \varepsilon - \mu\ (\varepsilon + \beta\ diag(diag(\varepsilon))) + o(|\varepsilon|). \tag{D.3}$$

Then, the first-order Taylor expansion of $vec(\mathbf{G}^{(n+1)})$ in the vicinity of the separation ($\varepsilon = \mathbf{0}$) is given by

$$vec(\mathbf{G}^{(n+1)}) = vec\,\mathbf{I} + (\mathscr{I} - \mu(\mathscr{I} + \beta\,diag(vec\,\mathbf{I}))\ vec(\mathbf{G}^{(n)} - \mathbf{I}), \tag{D.4}$$

where $\mathscr{I}$ is the $N^2 \times N^2$ identity matrix. Thus, the sought Jacobian is

$$\mathscr{J} = \mathscr{I} - \mu(\mathscr{I} + \beta\,diag(vec\,\mathbf{I})). \tag{D.5}$$

Appendix E. Local convergence of the GCII+ algorithm

The iteration of the GCII+ algorithm rewritten in terms of the global transfer matrix is given by

$$\mathbf{G}^{(n+1)} = \left(\mathbf{I} - \mu\left(\sum_{\beta\in\Omega^+} w_\beta\mathbf{C}_{y,y}^{1,\beta}\mathbf{S}_y^\beta - \mathbf{I}\right)\right)\mathbf{G}^{(n)}. \tag{E.1}$$

Let ε be again defined as a small perturbation at the separation solution such that $\mathbf{G}^{(n)} = \mathbf{I} + \varepsilon$. Here, for the GCII+ algorithm, we will assume that the scaling of the sources is such that at the separation, $\mathbf{G}^{(n)} = \mathbf{I}$, the following relation holds true:

$$\sum_{\beta\in\Omega^+} w_\beta\mathbf{C}_{s,s}^{1,\beta}\mathbf{S}_s^\beta = \mathbf{I}. \tag{E.2}$$

Then, the first-order Taylor expansion in ε of the iteration (E.1) is given by

$$\mathbf{G}^{(n+1)} = \mathbf{I} + (1 - \mu)\varepsilon - \mu w_1\mathbf{C}_{s,s}^{1,1}\varepsilon^{\mathrm{T}}$$

$$- \mu\left(\sum_{\beta\in\Omega,\beta\neq1} \beta w_\beta\mathbf{C}_{s,s}^{1,\beta}\mathbf{S}_y^\beta\right) diag(\varepsilon) + o(|\varepsilon|). \tag{E.3}$$

Similarly to previous algorithms, the local behavior of the iteration decouples for the diagonal and off-diagonal terms. For the diagonal terms ($G_{ii} = [\mathbf{G}]_{ii}$ with $i = 1, \ldots, N$) the iteration is

$$G_{ii}^{(n+1)} = 1 + \left(1 - \mu w_1 C_{s_i}^2 - \mu \sum_{\beta \in \Omega, \beta \neq 1} \beta w_\beta |C_{s_i}^{1+\beta}| \right) \varepsilon_{ii} + o(|\varepsilon|), \qquad (E.4)$$

whereas for the rest of terms $\forall i, j|_{i \neq j} = 1, \ldots, N$ is

$$\begin{pmatrix} G_{ij}^{(n+1)} \\ G_{ij}^{(n+1)} \end{pmatrix} = \begin{pmatrix} 1 - \mu & -\mu w_1 C_{s_i}^2 \\ -\mu w_1 C_{s_j}^2 & 1 - \mu \end{pmatrix} \begin{pmatrix} \varepsilon_{ij} \\ \varepsilon_{ji} \end{pmatrix} + o(|\varepsilon|). \qquad (E.5)$$

Using $\mathbf{g} = vec(\mathbf{G})$ and taking into account (E.4) and (E.5) it is easy to check that the Jacobian of the iteration $\mathscr{J} = \partial \mathbf{g}^{(n+1)}/\partial \mathbf{g}^{(n)}$ at the separation has the following eigenvalues $\forall i, j = 1, \ldots, N$:

$$\lambda_{i,i} = 1 - \mu \left(w_1 C_{s_i}^2 + \sum_{\beta \in \Omega} \beta w_\beta |C_{s_i}^{1+\beta}| \right), \qquad (E.6)$$

$$\lambda_{i,j} = 1 - \mu (1 \pm w_1 \sqrt{C_{s_i}^2 C_{s_j}^2}). \qquad (E.7)$$

These eigenvalues, in contrast to the previous case, depend on the source statistics (in particular they depend on the source's power). For non-i.i.d. sources, this fact results in different rates of convergence for the off-diagonal terms.

The local convergence is guaranteed whenever the modulus of all the eigenvalues is strictly < 1. Since at the separation $\sum_{\beta \in \Omega^+} w_\beta |C_{s_i}^{1+\beta}| = 1$ it is easy to deduce that $w_1 C_{s_i}^2 < 1 \ \forall i$ and therefore the term $w_1 \sqrt{C_{s_i}^2 C_{s_j}^2} < 1 \ \forall i, j$. Taking this fact into account we can see that the following step-size ensures the local convergence of the algorithm:

$$\mu < \min \left\{ \frac{2}{\max_{i=1,\ldots,N} \{ w_1 C_{s_i}^2 + \sum_{\beta \in \Omega} \beta w_\beta |C_{s_i}^{1+\beta}| \}}, \right.$$

$$\left. \frac{2}{1 + w_1 \max_{i=1,\ldots,N} \{ \sqrt{C_{s_i}^2 C_{s_j}^2} \}} \right\}. \qquad (E.8)$$

Simplifying the above condition by using $\max_{i=1,\ldots,N} \{ w_1 C_{s_i}^2 + \sum_{\beta \in \Omega} \beta w_\beta |C_{s_i}^{1+\beta}| \} \leqslant \beta_{\max}$ where $\beta_{\max} = \max_{\beta \in \Omega} \{ \beta \}$, and that $w_1 \max_{i=1,\ldots,N} \{ \sqrt{C_{s_i}^2 C_{s_j}^2} \} < 1$ we can conclude that these constraints are satisfied whenever we choose

$$\mu < \frac{2}{\beta_{\max}}. \qquad (E.9)$$

Appendix F. Local convergence of the CEASI algorithm

The CEASI algorithm can be rewritten in terms of the global transfer matrix as

$$\mathbf{G}^{(n+1)} = \left(\mathbf{I} - \mu_1(\mathbf{C}_{y,y}^{1,1} - \mathbf{I}) - \mu_2 \sum_{\beta \in \Omega} w_\beta (\mathbf{C}_{y,y}^{1,\beta} \mathbf{S}_y^\beta - \mathbf{S}_y^\beta \mathbf{C}_{y,y}^{\beta,1}) \right) \mathbf{G}^{(n)}. \tag{F.1}$$

Assuming now a unity variance scaling of the sources ($\mathbf{C}_{s,s}^{1,1} = \mathbf{I}$) and defining ε as a small perturbation of the separation solution such that $\mathbf{G}^{(n)} = \mathbf{I} + \varepsilon$, the Taylor expansion of iteration (F.1), at the separation, in terms of ε, is given by

$$\mathbf{G}^{(n+1)} = (\mathbf{I} + \varepsilon) - \mu_1(\varepsilon + \varepsilon^{\mathrm{T}}) - \mu_2 \sum_{\beta \in \Omega} w_\beta (\varepsilon \mathbf{C}_{s,s}^{1,\beta} \mathbf{S}_y^\beta - \mathbf{S}_y^\beta \mathbf{C}_{s,s}^{\beta,1} \varepsilon^{\mathrm{T}}) + o(|\varepsilon|). \tag{F.2}$$

Therefore, the adaptation of the diagonal terms of $\mathbf{G}$ for $i = 1, \ldots, N$ is

$$G_{ii}^{(n+1)} = 1 + (1 - 2\mu_1)\varepsilon_{ii} + o(|\varepsilon|), \tag{F.3}$$

whereas, for the remaining terms $\forall i,j|_{i \neq j} = 1, \ldots, N$ is

$$\begin{pmatrix} G_{ij}^{(n+1)} \\ G_{ij}^{(n+1)} \end{pmatrix} = \begin{pmatrix} 1 - \mu_1 - \mu_2 \displaystyle\sum_{\beta \in \Omega} w_\beta |C_{s_j}^{1+\beta}| & -\mu_1 + \mu_2 \displaystyle\sum_{\beta \in \Omega} w_\beta |C_{s_i}^{1+\beta}| \\ -\mu_1 + \mu_2 \displaystyle\sum_{\beta \in \Omega} w_\beta |C_{s_j}^{1+\beta}| & 1 - \mu_1 - \mu_2 \displaystyle\sum_{\beta \in \Omega} w_\beta |C_{s_i}^{1+\beta}| \end{pmatrix}$$

$$\times \begin{pmatrix} \varepsilon_{ij} \\ \varepsilon_{ji} \end{pmatrix} + o(|\varepsilon|). \tag{F.4}$$

It is easy to check that the eigenvalues of the Jacobian of the iteration $\mathscr{J} = \partial \mathbf{g}^{(n+1)} / \partial \mathbf{g}^{(n)}$ are given by

$$\lambda_{i,i} = 1 - 2\mu_1, \tag{F.5}$$

$$\lambda_{i,j} = \begin{cases} 1 - 2\mu_1 \\ 1 - \mu_2 \left(\displaystyle\sum_{\beta \in \Omega} w_\beta (|C_{s_i}^{1+\beta}| + |C_{s_j}^{1+\beta}|) \right) \end{cases} \tag{F.6}$$

$\forall i,j$. We can see from this result that the convergence depends on the $(1 + \beta)$-order cumulants of the signals. This fact will destroy the isotropic convergence property for non-i.i.d. sources.

Sufficient conditions for the local convergence are

$$0 < \mu_1 < 1, \tag{F.7}$$

$$0 < \mu_2 < 2 \left(\max_{i,j=1,\ldots,N} \left\{ \sum_{\beta \in \Omega} w_\beta (|C_{s_i}^{1+\beta}| + |C_{s_j}^{1+\beta}|) \right\} \right)^{-1}. \tag{F.8}$$

Inside this intervals we propose adaptation step-sizes that are simpler to evaluate and, at the same time, close to their optimal value for a quick convergence. These are

$$\mu_1 = \frac{1}{2}, \tag{F.9}$$

$$\mu_2 = \frac{1}{2} \left(\max_{i=1,\dots,N} \left\{ \sum_{\beta \in \Omega} w_\beta |C_{s_i}^{1+\beta}| \right\} \right)^{-1}. \tag{F.10}$$

Appendix G. Cumulants in terms of moments

In this last appendix we will see how to evaluate the cumulants of the outputs. This is necessary for the algorithm implementations. An easy way is to rewrite them in terms of the moments of the outputs by using the following formula (see [27]):

$$Cum(y_1, y_2, \dots, y_n)$$

$$= \sum_{(p_1,\dots,p_m)} (-1)^{m-1}(m-1)! \cdot E\left[\prod_{i \in p_1} y_i \right] E\left[\prod_{i \in p_2} y_i \right] \dots E\left[\prod_{i \in p_m} y_i \right], \tag{G.1}$$

where the sum is extended to all the possible partitions $(p_1, \dots, p_m)$, $m = 1, \dots, n$, of the set of natural numbers $(1, \dots, n)$.

This calculus results in simple complexity for lower orders but it quickly increases for higher orders. In our case, the fact that the cross-cumulants take the form $\mathbf{C}_{y,y}^{1,\beta}$ considerably simplifies this task for real and zero mean sources because many partitions disappear or give rise to the same kind of sets. Indeed, defining the moment $\mathbf{M}_y^\alpha = E[\mathbf{y}^{\odot \alpha}]$ and cross-moment matrices of the outputs as $\mathbf{M}_{y,y}^{1,\beta} = E[\mathbf{y}(\mathbf{y}^{\odot \beta})^\mathrm{T}]$, we can present below the expression of the cross-cumulant matrices $\mathbf{C}_{y,y}^{1,\beta}$ in terms of the moment matrices for $\beta = 1, \dots, 7$.

$$\mathbf{C}_{y,y}^{1,1} = \mathbf{M}_{y,y}^{1,1} = E[\mathbf{yy}^\mathrm{T}],$$

$$\mathbf{C}_{y,y}^{1,2} = \mathbf{M}_{y,y}^{1,2} = E[\mathbf{y}(\mathbf{y}^{\odot 2})^\mathrm{T}],$$

$$\mathbf{C}_{y,y}^{1,3} = \mathbf{M}_{y,y}^{1,3} - 3\mathbf{M}_{y,y}^{1,1} \, diag(\mathbf{M}_y^2),$$

$$\mathbf{C}_{y,y}^{1,4} = \mathbf{M}_{y,y}^{1,4} - 4\mathbf{M}_{y,y}^{1,1} \, diag(\mathbf{M}_y^3) - 6\mathbf{M}_{y,y}^{1,2} \, diag(\mathbf{M}_y^2),$$

$$\mathbf{C}_{y,y}^{1,5} = \mathbf{M}_{y,y}^{1,5} - 5\mathbf{M}_{y,y}^{1,1} \, diag(\mathbf{M}_y^4) - 10\mathbf{M}_{y,y}^{1,2} \, diag(\mathbf{M}_y^3) - 10\mathbf{M}_{y,y}^{1,3} \, diag(\mathbf{M}_y^2)$$

$$\qquad + 30\mathbf{M}_{y,y}^{1,1} \, diag(\mathbf{M}_y^2)^2,$$

$$\mathbf{C}_{y,y}^{1,6} = \mathbf{M}_{y,y}^{1,6} - 6\mathbf{M}_{y,y}^{1,1} \, diag(\mathbf{M}_y^5) - 15\mathbf{M}_{y,y}^{1,2} \, diag(\mathbf{M}_y^4) - 20\mathbf{M}_{y,y}^{1,3} \, diag(\mathbf{M}_y^3)$$

$$\qquad - 15\mathbf{M}_{y,y}^{1,4} \, diag(\mathbf{M}_y^2) + 120\mathbf{M}_{y,y}^{11} \, diag(\mathbf{M}_y^2) \, diag(\mathbf{M}_y^3)$$

$$\qquad + 90\mathbf{M}_{y,y}^{1,2} \, diag(\mathbf{M}_y^2)^2,$$

$$\mathbf{C}_{y,y}^{1,7} = \mathbf{M}_{y,y}^{1,7} - 7\mathbf{M}_{y,y}^{1,1} \, diag(\mathbf{M}_y^6) - 21\mathbf{M}_{y,y}^{1,2} \, diag(\mathbf{M}_y^5) - 35\mathbf{M}_{y,y}^{1,3} \, diag(\mathbf{M}_y^4)$$

$$- 35\mathbf{M}_{y,y}^{1,4} \, diag(\mathbf{M}_y^3) - 21\mathbf{M}_{y,y}^{1,5} \, diag(\mathbf{M}_y^2) + 210\mathbf{M}_{y,y}^{1,1} \, diag(\mathbf{M}_y^2) \, diag(\mathbf{M}_y^4)$$

$$+ 140\mathbf{M}_{y,y}^{1,1} \, diag(\mathbf{M}_y^3)^2 + 420\mathbf{M}_{y,y}^{1,2} \, diag(\mathbf{M}_y^2) \, diag(\mathbf{M}_y^3)$$

$$+ 210\mathbf{M}_{y,y}^{1,3} \, diag(\mathbf{M}_y^2)^2 - 630\mathbf{M}_{y,y}^{1,1} \, diag(\mathbf{M}_y^2)^3.$$

For the case of complex sources the expressions are much more complicated. As an example, we can see that for $\beta = 3$ the cumulant matrix is

$$\mathbf{C}_{y,y}^{1,3} = E[\mathbf{y}(\mathbf{y} \odot \mathbf{y}^* \odot \mathbf{y}^*)^{\mathrm{T}}] - E[\mathbf{y}\mathbf{y}^{\mathrm{T}}] \, diag(E[\mathbf{y}^* \odot \mathbf{y}^*])$$

$$- 2E[\mathbf{y}\mathbf{y}^{\mathrm{H}}] \, diag(E[\mathbf{y} \odot \mathbf{y}^*]). \tag{G.2}$$

Compare it with the real case and note the increase of complexity.

References

[1] L.B. Almeida, F.M. Silva, Adaptive decorrelation, Artificial Neural Networks, Elsevier, Amsterdam, Vol. 2, 1992, pp. 149–156.

[2] S. Amari, Natural gradient work efficiently in learning, Neural Comput. 10 (2) (1998) 251–276.

[3] S. Amari, Natural gradient learning for over- and under-complete bases in ica, Neural Comput. 11 (1999) 1875–1883.

[4] S. Amari, J.F. Cardoso, Blind source separation—semiparametric statistical approach, IEEE Trans. Signal Process. 45 (11) (1997) 2692–2697.

[5] S. Amari, A. Cichocki, Adaptive blind signal processing—neural network approaches, Proc. IEEE 86 (10) (1998) 2026–2048.

[6] A.J. Bell, T.J. Sejnowski, An information maximization approach to blind separation and blind deconvolution, Neural Comput. 7 (1996) 1129–1159.

[7] Brandwood, A complex gradient operator and its application in adaptive theory, IEE Proc. 130(1) (1983) 11–16.

[8] X.R. Cao, R.W. Liu, General approach to blind source separation, IEEE Trans. Signal Process. 44 (3) (1996) 562–571.

[9] J.F. Cardoso, Infomax and maximum likelihood for blind separation, IEEE Signal Process. Lett. 4 (4) (1997).

[10] J.F. Cardoso, Blind signal separation: statistical principles, Proc. IEEE 86 (10) (1998) 2009–2025.

[11] J.F. Cardoso, B. Laheld, Equivariant adaptive source separation, IEEE Trans. Signal Process. 44 (12) (1996) 3017–3030.

[12] J.-F. Cardoso, A. Souloumiac, Blind beamforming for non-Gaussian signals, IEE Proc.-F 140 (6) (1993) 362–370.

[13] A. Cichocki, R. Unbehauen, Robust neural networks with on-line learning for blind identification and blind separation of sources, IEEE Trans. Circuits Syst.-I 43 (11) (1996) 894–906.

[14] P. Comon, Independent component analysis, a new concept? Signal Process. 3 (36) (1994) 287–314.

[15] S. Cruces, Una visión unificada de los algoritmos de separacion ciega de fuentes (an unified view of blind source separation algorithms), Ph.D. Thesis, http://viento.us.es/sergio/PhD/thesis2.ps.gz, University of Vigo, Signal Processing Department, Spain, 1999.

[16] S. Cruces, A. Cichocki, L. Castedo, An iterative inversion approach to blind source separation, IEEE Trans. Neural Networks 11 (6) (2000) 1423–1437.

[17] M. Girolami, An alternative perspective on adaptive independent component analysis algorithms, Neural Comput. 10 (8) (1998) 2103–2114.

[18] A. Hyvarinen, E. Oja, A fast fixed-point algorithm for independent component analysis, Neural Comput. 9 (1997) 1483–1492.

[19] C. Jutten, J. Herault, Blind separation of sources. Part I: an adaptive algorithm based on neuromimetic architecture, Signal Process. 24 (1991) 1–10.

[20] J. Karhunen, E. Oja, L. Wang, R. Vigario, J. Koutsensalo, A class of neural networks for independent component analysis, IEEE Trans. Neural Networks 8 (3) (1997) 486–503.

[21] C.T. Kelley, Iterative methods for linear and nonlinear equations, in: Frontiers in Applied Mathematics, Vol. 16, SIAM, Philadelphia, PA, 1995, pp. 71–78.

[22] J.L. Lacoume, M. Gaeta, Source separation without a priori knowledge: the maximum likelihood solution, in: Torres, Masgrau, Lagunas, (Eds.), Proceedings of the EUSIPCO Conference, Barcelona, Elsevier, Amsterdam, 1990, pp. 621–624.

[23] R.H. Lambert, C.L. Nikias, Blind deconvolution of multipath mixtures, in: S. Haykin (Ed.), Unsupervised Adaptive Filtering, Vol. I, Wiley, New York, 2000.

[24] T.W. Lee, Independent Component Analysis: Theory and Applications, Kluwer Academic Publishers, Dordrecht, 1998, pp. 42–49.

[25] E.A. Lee, D.G. Messerschmitt, Digital Communication, Kluwer Academic Publishers, Dordrecht, 1990.

[26] E. Moreau, O. Macchi, High order contrasts for self-adaptive source separation criteria for complex source separation, Int. J. Adaptive Control Signal Process. 10 (1996) 19–46.

[27] C. Nikias, A. Petropulu, Higher Order Spectra Analysis: A Non-Linear Signal Processing Framework, Prentice-Hall, Englewood Cliffs, NJ, 1993, pp. 10–19.

[28] E. Oja, Nonlinear PCA criterion and maximum likelihood in independent component analysis, Proceedings of the First International Workshop on Independent Component Analysis (ICA'99), Aussois, France, January 1999, pp. 143–148.

[29] D.T. Pham, P. Garat, Blind separation of mixtures of independent sources through a quasi-maximum likelihood approach, IEEE Trans. Signal Process. 45 (7) (1997) 1712–1725.

[30] H.H. Yang, S. Amari, Adaptive on-line learning algorithms for blind source separation—maximum entropy and minimum mutual information, Neural Comput. (1997).

[31] V. Zarzoso, A.K. Nandi, Adaptive blind separation for virtually any source probability density function, IEEE Trans. Signal Process. 48 (2) (2000) 477–488.

Sergio A. Cruces-Alvarez was born in Vigo, Spain, in 1970. He received the Telecommunication Engineer degree in 1994 and the Ph.D. degree in 1999, both at the University of Vigo (Spain). From 1994 to 1995 he was working as a project engineer for the Department of Signal Theory and Communications of this university. In 1995 he joined the Signal Theory and Communications group of the University of Seville, where he is currently an Associate Professor. He teaches undergraduate and graduate courses on digital signal processing, adaptive filter theory and mathematical methods for communication. In 1997 and 1999 he was invited to visit the Laboratory for Advanced Brain Signal Processing under the Frontier Research Program RIKEN (Japan). His current research interests include statistical signal processing (especially blind source separation and independent component analysis), information theoretic and neural network approaches, blind equalization and filter stabilization techniques.

Luis Castedo-Ribas was born in Santiago de Compostela, Spain, in 1966. He received the Ingeniero de Telecomunicación and Doctor Ingeniero de Telecomunicación degrees, both from Universidad Politécnica de Madrid (UPM), Spain, in 1990 and 1993, respectively. From 1990 to 1994 he was with the Departamento de Señales, Sistemas y Radiocomunicación at the UPM where he worked in array processing applied to digital communications. In 1994, he joined the Departamento de Electrónica y Sistemas at Universidad de A Coruña, Spain, where he is currently Associate Professor and teaches courses in signal processing, digital communications and linear control systems. His research interests include blind adaptive filtering and signal processing methods for space and code diversity exploitation in communication systems.

Andrzej Cichocki received the M.Sc. (with honors), Ph.D., and Habilitate Doctorate (Dr.Sc.) degrees, all in electrical engineering, from Warsaw University of Technology (Poland) in 1972, 1975, and 1982, respectively. Since 1972, he has been with the Institute of Theory of Electrical Engineering and Electrical Measurements at the Warsaw University of Technology, where he became a full Professor in 1991. He is

the co-author of two books: MOS Switched-Capacitor and Continuous-Time Integrated Circuits and Systems (Springer-Verlag, 1989) and Neural Networks for Optimization and Signal Processing (Teubner-Wiley, 1993/94) and more than 150 research papers. He spent a few years at University Erlangen–Nuernberg (GERMANY) as Alexander Humboldt Research Fellow and Guest Professor. Since 1995 he has been working in the Brain Science Institute RIKEN (Japan), as a team leader of the Laboratory for Open Information Systems and currently as a head of laboratory for Advanced Brain Signal Processing. His current research interests include neural networks, biomedical signal and image processing, especially analysis and processing of multi-sensory EEG/MEG, PET, and microphone array data. He spent 3 yr as a member of the Technical Committee for Neural Network for Signal Processing and recently was a member of core group who established a new IEEE Circuits and Systems Technical Committee for Blind Signal Processing. More information about his research activities can be found at http://www.bsp.brain.riken.go.jp/.