From Blind Signal Extraction to Blind Instantaneous Signal Separation: Criteria, Algorithms, and Stability

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

***Abstract*—This paper reports a study on the problem of the blind simultaneous extraction of specific groups of independent components from a linear mixture. This paper first presents a general overview and unification of several information theoretic criteria for the extraction of a single independent component. Then, our contribution fills the theoretical gap that exists between extraction and separation by presenting tools that extend these criteria to allow the simultaneous blind extraction of subsets with an arbitrary number of independent components. In addition, we analyze a family of learning algorithms based on Stiefel manifolds and the natural gradient ascent, present the nonlinear optimal activations (score) functions, and provide new or extended local stability conditions. Finally, we illustrate the performance and features of the proposed approach by computer-simulation experiments.**

***Index Terms*—Blind-signal extraction, blind signal separation, independent component analysis, negentropy and minimum entropy, projection pursuit.**

I. Introduction

THE problem of blind-signal extraction (BSE) consists of the recovery or estimation of part of the non-Gaussian independent components that appear linearly combined in the observations. Blind signal separation (BSS) is a special case of BSE in which one considers the simultaneous recovery of all the independent components from the observations. These problems form part of independent component analysis (ICA), an active field of research that has attracted great interest because of its large number of applications in diverse fields [1], [2].

The criteria to solve ICA problems are usually mathematically expressed in the form of the optimization of a function with some specific properties. These functions have a long history and different origins. In the late 1970s several *objective functions* (like kurtosis and standardized negative Shannon entropy) were proposed by geophysicists to solve the problem of blind deconvolution [3]–[7]. In the 1980s, part of this work evolved into a field of statistics named projection pursuit, which was concerned with finding interesting low-dimensional informative views of high-dimensional data sets [8]–[11]. These projections were automatically obtained by maximizing some indexes of interest (the standardized absolute cumulants, the exponential Shannon entropy, Fisher information, etc.). It was nearly at this time, after the pioneering work of Jutten and Herault [12], when the field of ICA was created (see [13], [14] and references therein), stressing the importance of the Darmois–Skitovitch theorem [15], [16] and proposing new *information theoretic contrasts* as driven criteria to solve the BSS problem [15]–[30].

ICA can be computationally very demanding if the number of source signals is large (say, on the order of 100 or more). In particular, this is the case in biomedical signal-processing applications such as electroencephalogram/magnetoencephalogram (EEG/MEG) data processing in which the number of sensors (observations) can be larger than 120 and it is desired to extract only some "interesting" components. Fortunately, BSE overcomes this difficulty by attempting to recover only a small subset of desired independent components from a large number of sensor signals. However, most of the existing BSE criteria and associated algorithms only recover one independent component at a time, or all at the same time (in the case of BSS). The sequential extraction of several components from the mixture is obtained by alternating BSE with Gaussianization of the extracted components [11] or with their deflation [33].

The objectives of this paper are twofold. One is to provide evidence that the projection pursuit methodology provides a unified framework for the different criteria that can solve ICA problems. The second is to show that the standard approach for separated treatment of BSS and BSE is somewhat artificial, since there exists general and unified criteria that can be used in both cases. We recently outlined briefly in a letter [34] how to extend the classical criteria for BSS and BSE of one independent component to the case of simultaneous blind source extraction of an arbitrary subgroup of P $(1 \leq P \leq N)$ independent components, where P is specified by the user. In this paper, we further develop this approach and complete it with full proofs of the results.

The structure of the paper is as follows. Section II specifies the considered signal model and notation. Section III illustrates the difficulties in the direct extension of some criteria from BSS to BSE. Section IV discusses and provides an overview of the existing contrast functions that are suitable for extraction of a single independent component, and Section V introduces the tools that extend these functions to allow simultaneous extraction of arbitrary subsets of independent components. Section VI

Manuscript received March 15, 2003; revised October 27, 2003.

S. A. Cruces-Alvarez is with the Area de Teoría de la Señal y Comunicaciones, Escuela de Ingenieros, Universidad de Sevilla, Camino de los descubrimientos s/n, 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 Electrical Engineering Department, Warsaw University of Technology, 00-661 Warsaw, Poland (e-mail: cia@bsp.brain.riken.go.jp).

S. Amari is with the Brain-Style Information Systems Research Group, Brain Science Institute, Riken, 351-01 Saitama, Japan (e-mail: amari@brain.riken.go.jp).

Digital Object Identifier 10.1109/TNN.2004.828764

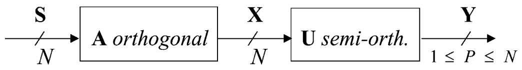

Fig. 1. Considered signal model for simultaneous BSE.

analyzes BSE criteria that take into account the additional information such as the probability density functions (pdf) of the desired sources. In Sections VII and VIII, we consider the use of the natural gradient on the Stiefel manifold to perform the constrained optimization of the specific contrast functions. We also present practical upper bounds for the step size of the algorithm derived from the asymptotical-stability analysis. These bounds allow us to dramatically improve the convergence speed of the learning algorithm. Section IX discusses exemplary simulation results and finally, Section X presents the conclusions.

II. Signal Model and Notation

Let us consider the signal model of Fig.1, where N unknown, statistically independent source signals, usually called sources or components are drawn from a random vector process $\mathbf{S}(t) = [S_1(t), \cdots, S_N(t)]^T$. The sources are linearly mixed in the memoryless system, described by a nonsingular mixing matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, to give the vector of observations

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

It is commonly assumed that the vector of sources $\mathbf{S}(t)$ has a mean of zero and a normalized-covariance matrix ($\mathrm{E}[\mathbf{S}(t)] = \mathbf{0}$, $Cov(\mathbf{S}(t)) = \mathbf{I}_N$), where $\mathrm{E}[\cdot]$ denotes the expectation operator and $\mathbf{I}_N$ the identity matrix of dimension $N \times N$. Without loss of generality, we consider that the unknown mixing matrix $\mathbf{A}$ is orthogonal

$$\mathbf{A}\mathbf{A}^T = \mathbf{I}_N \tag{2}$$

because the orthogonality of the mixing matrix can be always enforced by performing prewhitening on the original observations. For the sake of mathematical simplicity with some of the studied criteria, we will not address, in this paper, the case of noisy data. From hereafter, since the mixing system is memoryless, we will drop the time index when referring to the random variables of the considered processes. Under these hypotheses, the determinant of the mixing system simplifies to unity, $|\mathbf{A}| = 1$, and the joint density of the observations is coincident with the product of marginal densities of the sources

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{p_{\mathbf{S}}(\mathbf{s})}{|\mathbf{A}|} = \prod_{i=1}^{N} p_{S_i}(s_i). \tag{3}$$

Under the linear mixing model (1), the Darmois–Skitovitch theorem [16] guarantees the identifiability of the original non-Gaussian sources from the observations up to a permutation and scaling of them. Then, in order to extract P non-Gaussian sources from the mixture (where $1 \leq P \leq N$), the observations will be processed by a linear and memoryless extracting system characterized by an $P \times N$ semiorthogonal matrix $\mathbf{U}$ satisfying $\mathbf{U}\mathbf{U}^T = \mathbf{I}_P$. This yields the vector process of outputs or estimated sources

$$\mathbf{Y} = \mathbf{U}\mathbf{X} = \mathbf{G}\mathbf{S} \tag{4}$$

where $\mathbf{G} = \mathbf{U}\mathbf{A}$ is the global transfer matrix (of dimensions $P \times N$) from the sources to the outputs. The semiorthogonality of the extracting and global transfer matrices will be important for preserving the spatial decorrelation of the outputs since $\mathrm{Cov}(\mathbf{Y}) = \mathbf{G}\mathbf{G}^T = \mathbf{I}_P$.

In this paper, we adopt the following notation. We work with normalized random variables (those with zero mean and unit variance) and, according to the standard notations, we employ capital letters for random variables and lowercase letters for the samples of these variables. For any given random variable Y with mean $\mathrm{E}[Y]$ and covariance $\mathrm{Cov}(Y)$, the notation $Y^{\mathcal{N}}$ specifies a normal (Gaussian) random variable with the same mean and covariance. Similarly, $\mathbf{Y}^{\mathcal{N}}$ represents a Gaussian random vector with the same mean and covariance matrix as the random vector $\mathbf{Y}$. The rth order autocumulant of the random variable Y is denoted by $C_Y^r = \mathrm{Cum}(Y \times r)$. The joint differential entropy of the random vector $\mathbf{Y}$ with density $p_{\mathbf{Y}}$ is expressed as

$$h(\mathbf{Y}) = -\int p_{\mathbf{Y}}(\mathbf{y}) \log p_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y} \tag{5}$$

with the convention that $0 \log 0 = 0$. We assume that the log operator denotes the natural logarithm, so the entropy will be specified in *nats*. For a Gaussian random vector of dimension P with uncorrelated and normalized components, we have $h(\mathbf{Y}^{\mathcal{N}}) = P/2 \log(2\pi e)$. The Kullback–Leibler divergence, or *relative entropy* between two continuous multivariate densities $p_{\mathbf{R}}(\mathbf{y})$ and $p_{\mathbf{V}}(\mathbf{y})$, of the same dimension, is defined as

$$D(p_{\mathbf{R}} \| p_{\mathbf{V}}) = \int p_{\mathbf{R}}(\mathbf{y}) \log \frac{p_{\mathbf{R}}(\mathbf{y})}{p_{\mathbf{V}}(\mathbf{y})} d\mathbf{y}. \tag{6}$$

This divergence is nonnegative $D(p_{\mathbf{R}} \| p_{\mathbf{V}}) \geq 0$, with equality only if $p_{\mathbf{R}}(\mathbf{y}) = p_{\mathbf{V}}(\mathbf{y})$ almost everywhere [31].

Sometimes, it will be useful to complete the output vector $\mathbf{Y} = [Y_1, \ldots, Y_P]^T$ from dimension P to dimension N. This is done by defining a complementary output vector of random variables $\mathbf{Y}_c = [\tilde{Y}_{P+1}, \ldots, \tilde{Y}_N]^T$ orthogonal to $\mathbf{Y}$, and grouping them together in a new virtual-output vector process $\tilde{\mathbf{Y}} = [\mathbf{Y}^T, \mathbf{Y}_c^T]^T$ whose covariance is the identity matrix, i.e. $\mathrm{Cov}(\tilde{\mathbf{Y}}) = \mathbf{I}_N$.

III. Does the Natural Criterion for BSS Extend to BSE?

The most natural criterion for BSS of N sources is based on the minimization of the mutual information of the outputs

$$\begin{aligned} I(Y_1, \ldots, Y_N) &= D\left(p_{\mathbf{Y}} \| \prod_{i=1}^{N} p_{Y_i}\right) \\ &= -h(Y_1, \ldots, Y_N) + \sum_{i=1}^{N} h(Y_i) \end{aligned} \tag{7}$$

since the independence of the sources and the non-Gaussianity of at least $N-1$ of them are the key assumptions in this problem [15]. This was the starting approach for several interesting BSS algorithms. The main difficulty in applying this criterion is the necessity to estimate the joint entropy of the outputs, since this involves the estimation of their joint pdf, a nontrivial task that would require an extensive amount of data and computational resources. However, if the observations are prewhitened and under the spatial decorrelation constraint of the outputs, the joint entropy is kept constant $h(Y_1,\ldots,Y_N)=h(X_1,\ldots,X_N)$ and the criterion can now be implemented.

Unfortunately, the minimum mutual-information criterion does not extend directly to the blind extraction of $P<N$ signals. On the contrary to the BSS case, having P-independent outputs does not necessarily correspond to the extraction of one independent component at each output. At the minima of

$$I(Y_1,\ldots,Y_P)=D\left(p_{\mathbf{Y}}\|\prod_{i=1}^{P}p_{Y_i}\right) \tag{8}$$

the N sources split into P-disjoint groups and each of the independent outputs still can be a mixture of sources within the same group.

Since the direct extension seems to fail, a question arises: What are the suitable criteria for BSE? When trying to answer this question in this paper, one of our results will suggest that the maximization of the following contrast function:

$$\Psi_{M\mathrm{Neg}}(\mathbf{Y})=\sum_{i=1}^{P}D\left(p_{Y_i}\|p_{Y_i^{\mathcal{N}}}\right) \tag{9}$$

is a quite natural criterion for the extraction of subsets of P sources at the outputs. As we will show in forthcoming sections, the justification of this result has its roots, in part, in the methodology of *projection pursuit density estimation* (PPDE) of the observations proposed by Friedman *et al.* [9]. PPDE is a parametric technique that estimates the multivariate density of the observations $p_{\mathbf{X}}(\mathbf{x})$ and has the special feature that the search can be carried out in a low-dimensional setting (usually univariate), trying thereby to avoid the curse of dimensionality.

In PPDE, one first chooses an initial density estimate $p_{\hat{\mathbf{X}}}^{(0)}$ that reflects all the *a priori* knowledge of the data and that is maximally noncommittal with regard to the missing information. Let $\mathbf{U}_{i:}$ denote the ith row of the separating (or extracting) matrix $\mathbf{U}$; the method iteratively constructs factorial improved estimates of the form

$$p_{\hat{\mathbf{X}}}^{(i)}(\mathbf{x})=p_{\hat{\mathbf{X}}}^{(i-1)}(\mathbf{x})f_i(y_i) \tag{10}$$

where the augmenting function $f_i(\cdot)$ and the one-dimensional projection of the observations $y_i=\mathbf{U}_{i:}\mathbf{x}$ are chosen in such a way that they maximize the relative increment in the fit

$$\psi_{\mathrm{PPDE}}(\mathbf{U}_{i:},f_i)=D\left(p_{\mathbf{X}}\|p_{\hat{\mathbf{X}}}^{(i-1)}\right)-D\left(p_{\mathbf{X}}\|p_{\hat{\mathbf{X}}}^{(i)}\right). \tag{11}$$

IV. Extraction of a Single Source

When recovering a single independent component, the extracting system $\mathbf{U}$ will be a row vector of unit norm and there will be a single–output that we will denote as Y_1. In this section, we will assume a completely blind scenario where one knows only the observations and the existence of at least one non-Gaussian independent component in the mixture. However, there is no *a priori* information about the mixing matrix nor about the density of the desired source.

A. Contrast Functions and Indexes of Interest

To overcome the difficulties associated with the high dimensionality of the joint densities, we avoid working explicitly with them. With this aim, Huber suggested in [10] to find a functional $Q(\cdot)$ that maps the pdf of the output random variable Y_1 to a real index $Q(Y_1)$ which satisfies

$$Q(Y_1)\le\max\{Q(S_1),\ldots,Q(S_N)\} \tag{12}$$

with equality only if Y_1 is, at least, equivalent (in some given sense) to one of the independent components. Therefore, a proper optimization of $Q(Y_1)$ will lead to the extraction of one independent component.

Since the affine transformations of the one-dimensional random variable Y_1 are not relevant to this problem, a good class of functionals $Q(\cdot)$ are those that are invariant for all the members of the equivalence class defined by the affine transformations of the argument, i.e., the ones that satisfy $Q(\alpha Y_1+\beta)=Q(Y_1)\forall\alpha\neq 0$, α, $\beta\in\mathbb{R}$. An alternative approach consists of using semiorthogonal indexes $Q(\cdot)$. A semiorthogonal index will constrain the argument of the functional to be a normalized random variable (of zero mean and unit variance), avoiding, in this way, the affine invariance requirement.

The properties of the Huber's indexes $Q(\cdot)$ are similar and related to the properties of contrast functions introduced independently by Donoho [7] and by Comon [15]. Indexes and contrasts represent, in fact, equivalent concepts.

Some of the results of this paper will apply to the class of the semiorthogonal contrast functions $\{\psi(\cdot)\}$ that impose the constraint on the outputs to be normalized random variables and that possess the following important properties:

Property 1: The contrast function $\psi(\cdot)$ satisfies a *weak form of strict convexity*, which is defined only with respect to the linear combinations of the independent components, in the sense that, if $Y_1=\sum_{j=1}^{N}G_{1j}S_j$ and $\sum_{j=1}^{N}|G_{1j}|^2=1$, then

$$\psi(Y_1)\le\sum_{j=1}^{N}|G_{1j}|^2\psi(S_j) \tag{13}$$

where, for $\psi(Y_1)>0$, the equality holds true if and only if one of the independent components is extracted.

Property 2: The contrast function $\psi(\cdot)$ is always nonnegative and by convention, we can assign to $\psi(\cdot)$ the value zero when the argument is a random variable with Gaussian distribution.

These properties are consistent with the idea of assigning to the independent components the local maxima of the index of interest over certain subspaces (property 1) and to the Gaussian distribution the least interesting index (property 2). More specifically, let us consider an ordering of the independent components such that $\psi(S_1)\ge\psi(S_2)\ge\ldots\ge\psi(S_N)$. Properties 1 and 2 imply that the ith source maximizes $\psi(Y)$ over the subspace spanned by the linear combinations of the independent components that range from i to N. Consequently, whenever they hold true, S_1 will be a global maximum of the index of interest.

In the following sections, we will see that most of the known information theoretic criteria for the blind extraction of a single source can be unified, interpreted, and represented in terms of an approximation to the density of the observations (the projection-pursuit density-estimation methodology). Some of the criteria are shown to be equivalent in the sense that they lead to the same contrast function.

B. The Negentropy and Minimum-Entropy Criteria

One of the key assumptions in the ICA/BSE problem is that the sources are mutually independent. Another two practical simplifying assumptions are that $\mathrm{E}[\mathbf{X}] = \mathbf{0}$ and $\mathrm{Cov}(\mathbf{X}) = \mathbf{I}_N$. The most reasonable initial approximation $p^{(0)}_{\mathbf{X}}$ of the density of the observations $p_{\mathbf{X}}(\mathbf{x})$ is the less biased estimate on the basis of the previous information [32]. The density estimate that is maximally noncommittal with regard to the missing information, is given by the N-dimensional normalized-Gaussian density $p_{\mathbf{X}^{\mathcal{N}}}(\mathbf{x}) = (2\pi)^{-N/2}\exp(-\mathbf{x}^T\mathbf{x}/2)$, which maximizes the differential entropy $h(\mathbf{X})$ among all the distributions with the given mean and covariance [31]. Note that this initial choice corresponds to the factorial decomposition of the density of the observations in N-independent Gaussian marginals

$$p_{\mathbf{X}^{\mathcal{N}}}(\mathbf{x}) = \prod_{i=1}^{N} p_{\tilde{Y}_{i^{\mathcal{N}}}}(\tilde{y}_i) \tag{14}$$

for any arbitrary orthogonal transformation $\tilde{\mathbf{y}} = \tilde{\mathbf{U}}\mathbf{x}$.

After the initial approximation has been chosen, PPDE tries to improve the fit of the estimate to the true density of the observations by means of a new multiplicative factor or augmenting function $f_1(y_1)$, which incorporates additional information from the output. The new estimate $p^{(1)}_{\hat{\mathbf{X}}}(\mathbf{x}) = p^{(0)}_{\hat{\mathbf{X}}}(\mathbf{x})f_1(\mathbf{U}\mathbf{x})$ is then optimized by maximizing the index (11) with respect to the functional $f_1(\cdot)$, while keeping the initial constraints. The projection-pursuit index simplifies for this case to

$$\begin{aligned}\psi_{\mathrm{PPDE}}(\mathbf{U}, f_1) &\\ &= D\left(p_{\tilde{\mathbf{Y}}}\|p_{\tilde{\mathbf{Y}}^{\mathcal{N}}}\right) - D\left(p_{\tilde{\mathbf{Y}}}\|f_1 p_{\tilde{\mathbf{Y}}^{\mathcal{N}}}\right) &(15)\\ &= D\left(p_{Y_1}\|p_{Y_1^{\mathcal{N}}}\right) - D\left(p_{Y_1}\|f_1 p_{Y_1^{\mathcal{N}}}\right) &(16)\end{aligned}$$

where the second divergence term at the right (which accounts for all the dependence with $f_1(y_1)$) is nonnegative and equal to zero only for the optimal augmenting function [9]

$$f_1^*(y_1) = \frac{p_{Y_1}(y_1)}{p_{Y_1^N}(y_1)} \tag{17}$$

for which one replaces the initial Gaussian marginal of the output y_1 in (14) by its corresponding true density, resulting the negentropy-density estimate of the observations

$$p^{(\mathrm{Neg})}_{\hat{\mathbf{X}}}(\mathbf{x}) = p_{Y_1}(y_1)p_{\mathbf{Y}_c^{\mathcal{N}}|Y_1^{\mathcal{N}}}(\mathbf{y}_c|y_1). \tag{18}$$

The optimized index with respect to the functional form of the factor f_1 simplifies to

$$\psi_{\mathrm{Neg}}(Y_1) = \psi_{\mathrm{PPDE}}\left(\mathbf{U}, f_1^*\right) = D\left(p_{Y_1}\|p_{Y_1^{\mathcal{N}}}\right) \tag{19}$$

and after the fit, the projection pursuit density estimation criterion reduces to find the density of the output Y_1 or, equivalently, the vector $\mathbf{U}$, which maximizes the divergence from the Gaussian density. This is usually known as the *negentropy* criterion [2], [15] in ICA. However, its origins are much older and can be traced back to the *minimum entropy*-criterion proposed by Godfrey [4] and Donoho [7] in single-channel blind deconvolution. This later criterion consists of finding the linear projection of the observations that minimizes the output entropy subject to a constant covariance constraint

$$\min h(Y_1) \quad \text{subject to} \quad \mathrm{Cov}(Y_1) = 1. \tag{20}$$

This is equivalent to maximizing the Negentropy contrast function

$$\psi_{\mathrm{Neg}}(Y_1) = \frac{1}{2}\log(2\pi e) - h(Y_1). \tag{21}$$

Thus, both criteria (*minimum entropy* and *negentropy*) are equivalent and can be interpreted in terms of a fitting to the observations density.

The following lemma (from [31]) provides a lower bound on the differential entropy of a sum of two independent random variables in terms of their individual differential entropies.

Lemma 1 (Entropy Power Inequality): If A and B are independent continuous random variables, then

$$e^{2h(A+B)} \geq e^{2h(A)} + e^{2h(B)} \tag{22}$$

with equality, if and only if A, B are Gaussian.

Using this lemma, we prove in Appendix that the *negentropy* index satisfies properties 1 and 2. Therefore

$$\psi_{\mathrm{Neg}}(Y_1) \leq \sum_{j=1}^{N} |G_{1j}|^2 \psi_{\mathrm{Neg}}(S_j) \tag{23}$$

with equality only when Y_1 is one of the independent components, i.e., it will be one of the sources if $\psi_{\mathrm{Neg}}(Y_1)$ is positive or an arbitrary linear combination of Gaussian sources if $\psi_{\mathrm{Neg}}(Y_1) = 0$. Thus, the global maximum of the contrast function is only achieved at the extraction of the independent component with smallest differential entropy (let us assume that this component is S_1). Accordingly, the Negentropy criterion seems to guide us to a form of Occam's razor principle: the best fit is obtained for $p^{(\mathrm{Neg})}_{\hat{\mathbf{X}}}(\mathbf{x}) = p_{S_1}(s_1)p_{\mathbf{Y}_c^{\mathcal{N}}|S_1^{\mathcal{N}}}(\mathbf{y}_c|s_1)$, when one chooses the simplest, the least uninformative or most structured projection $S_1 = \mathbf{U}\mathbf{X}$ to explain the density of the observations.

C. Divergence From the Uniform Density

Let us denote with $F(\cdot)$ the standard Gaussian distribution function. The transformation $Z_1 = F(Y_1)$ maps the Gaussian random variable $Y_1^{\mathcal{N}}$ to a uniform random variable $\mathcal{U} = F(Y_1^{\mathcal{N}})$, with bounded support in [0, 1].

The divergence from the uniform density is a criterion originally proposed by Claerbout [5] and later, independently, by Friedman *et al.* [11]. This criterion is based on the maximization of the relative entropy of a transformed density of the output p_{Z_1} with respect to the uniform density $p_{\mathcal{U}}$

$$\psi_{DU}(Y_1) = D\left(p_{Z_1}\|p_{\mathcal{U}}\right) = -h\left(F(Y_1)\right). \tag{24}$$

Again, the criterion chooses the simplest model in the sense of minimizing the differential entropy of the transformed output.

For this reason, Claerbout called it *minimum information deconvolution*.

Since the relative entropy is invariant under nonlinear invertible transformations $F(\cdot)$, we have that $D(p_{Z_1}\|p_{\mathcal{U}}) = D(p_{Y_1}\|p_{Y_1^{\mathcal{N}}})$, with the result that the divergence from the uniform density is just an alternative expression for the *negentropy* index

$$\psi_{D\mathcal{U}}(Y_1) = \psi_{\mathrm{Neg}}(Y_1). \tag{25}$$

D. Cumulants Based Indexes

The indexes based on cumulants have a long history and several authors have proposed them in many different ways and forms [3], [15], [27], [37]. The *higher order cumulants* of the outputs can be used in data corrupted by additive Gaussian noise because they are asymptotically invariant to the presence of such noise in the mixture. For finite samples this result is only approximate and usually does not hold for signal-to-noise ratios (SNRs) that are too low.

One general form of the cumulant based index, for normalized random variables, is given by

$$\psi_{\mathrm{Cum}}(Y_1) = \sum_{r>2} \omega_r' \left|C_{Y_1}^r\right|^{\alpha_r} \tag{26}$$

where $|C_{Y_1}^r|$ denotes the modulo of the rth-order autocumulant, $\omega_r' = \omega_r/(r\alpha_r)$ are nonnegative weighting factors such that $\sum_r \omega_r' = 1$, and the exponents α_r are greater than or equal to unity. Typically, $\alpha_r = 1$ when only one cumulant order is involved and $\alpha_r = 2$ if a set of different cumulants are to be jointly maximized.

The cumulant index also satisfies Properties 1 and 2 (see subsection B of the Appendix). Therefore

$$\psi_{\mathrm{Cum}}(Y_1) \leq \sum_{j=1}^{N} |G_{1j}|^2 \psi_{\mathrm{Cum}}(S_j) \tag{27}$$

with equality only for the extraction of one of the independent components $(Y_1 = S_i)$, provided that $\psi_{\mathrm{Cum}}(Y_1) \neq 0$. Note that for the Gaussian distribution $\psi_{\mathrm{Cum}}(Y_1^{\mathcal{N}}) = 0$, since the higher-order cumulants of Gaussian processes are all zero.

V. Simultaneous Extraction of P of the N Sources

As we have seen in the previous section, the blind extraction of one of the non-Gaussian sources is obtained by solving the following constrained maximization problem:

$$\max_{\mathbf{U}} \psi(Y_1) \quad \text{subject to} \quad \mathrm{Cov}(Y_1) = 1. \tag{28}$$

It is well known, however, that the BSS of the set of all the N sources (being at most one Gaussian) is obtained by maximizing

$$\max_{\mathbf{U}} \sum_{i=1}^{N} \psi(Y_i) \quad \text{subject to} \quad \mathrm{Cov}(\mathbf{Y}) = \mathbf{I}_N. \tag{29}$$

In this section, we will try to fill the theoretical gap between both previous approaches, showing that there is a continuum of contrasts of the form

$$\max_{\mathbf{U}} \sum_{i=1}^{P} \psi(Y_i) \quad \text{subject to} \quad \mathrm{Cov}(\mathbf{Y}) = \mathbf{I}_P. \tag{30}$$

Such a contrast function is suitable for the whole range of subproblems (from the extraction of a single independent component to the simultaneous extraction of all the independent components $(1 \leq P \leq N)$), and only involve the use of univariate densities. The following theorem, proved in Appendix, establishes a fundamental result for any contrast function $\psi(Y_i)$ satisfying properties 1 and 2.

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

$$\psi(S_1) \geq \ldots \geq \psi(S_P) > \psi(S_{P+1}) \geq \ldots \geq \psi(S_N) \tag{31}$$

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

$$\Psi(\mathbf{Y}) = \sum_{i=1}^{P} d_i \psi(Y_i) \quad \text{subject to} \quad \mathrm{Cov}(\mathbf{Y}) = \mathbf{I}_P \tag{32}$$

will be a contrast function whose global maxima correspond to the extraction of the first P sources from the mixture. If, additionally, $\psi(S_1) > \ldots > \psi(S_P)$ and $d_1 > d_2 > \ldots > d_P$, then the global maximum is unique and corresponds to the ordered extraction of the first P sources of the mixture, i.e., the global maximum is $\mathbf{Y} = [S_1, \ldots, S_P]^T$.

If we do not need to extract the components in any specific order, we can simply set $d_1 = d_2 = \ldots = d_P = 1$ and obtain the following special cases:

1) The cumulant contrast function for extraction of P of the N sources, with largest cumulant index, is

$$\Psi_{\mathrm{Cum}}(\mathbf{Y}) = \sum_{i=1}^{P} \sum_{r>2} \omega_r' \left|C_{Y_i}^r\right|^{\alpha_r} \text{ subject to } \mathrm{Cov}(\mathbf{Y}) = \mathbf{I}_P. \tag{33}$$

2) The *marginal negentropy* contrast function for the extraction of the P of the N sources, with minor entropy, is

$$\begin{aligned}\Psi_{M\mathrm{Neg}}(\mathbf{Y}) &= \sum_{i=1}^{P} D\left(p_{Y_i}\|p_{Y_i^{\mathcal{N}}}\right) \\ &= \frac{P}{2}\log(2\pi e) - \sum_{i=1}^{P} h(Y_i) \\ &\quad \text{subject to} \quad \mathrm{Cov}(\mathbf{Y}) = \mathbf{I}_P.\end{aligned} \tag{34}$$

Remark: We call this a *marginal negentropy* contrast function to distinguish it from the *negentropy* function

$$\Psi_{\mathrm{Neg}}(\mathbf{Y}) = D\left(p_{\mathbf{Y}}\|p_{\mathbf{Y}^{\mathcal{N}}}\right) \tag{35}$$

which results from applying the PPDE approach when considering multivariate statistics. Note that for $P = 1$, both coincide $(\Psi_{M\mathrm{Neg}}(Y_1) = \Psi_{\mathrm{Neg}}(Y_1))$, but for $P > 1$, the *negentropy* function $\Psi_{\text{Neg}}(\mathbf{Y})$ is not a valid contrast function for the extraction of P sources. This is because it is maximized for all the output vectors that belong to the subspace spanned by the P sources of lowest entropy. However, subtracting from it the mutual information of the outputs, one obtains the *marginal negentropy* contrast function, which results from the application of the theorem

$$\Psi_{M\text{Neg}}(\mathbf{Y}) = \Psi_{\text{Neg}}(\mathbf{Y}) - I(Y_1, \dots, Y_P). \tag{36}$$

There is a very interesting interpretation of the role played by each term of this index. The first term on the right acts as a preprocessing step that removes $N-P$ uninteresting sources by reducing the original nonsquare BSE problem into a $P \times P$ BSS problem. The second term on the right is the minimum mutual-information contrast function that will eventually solve the resulting BSS problem.

VI. The Partially Blind ICA Problem

Now we consider a partially blind or semiblind scenario in which the densities of the interesting or desired sources are assumed to be known and are non-Gaussian. The *maximum likelihood* and *infomax* criteria can incorporate relatively easily this additional information in the BSS case $(P = N)$. In this section, we will extend these criteria to the case of BSE $(P \le N)$. To allow us to identify the set of P-desired sources from the set of their densities, we will assume that there is a one-to-one correspondence between both sets, i.e., that the P-desired densities also determine the set of P-desired sources.

A. Maximum Likelihood

The *maximum likelihood* (ML) criterion for BSS is very popular in the ICA research community because its optimization depends only on the score functions of the densities of the sources (see [23] and [29]). In the following, we present the extension of this criterion to BSE. It should be noted that despite the similar form of the resulting ML criterion for BSE, optimization is much more difficult to perform.

When one knows *a priori*, the probability density function of the desired sources $p_{\mathcal{S}_i}(y_i)$, $i = 1, \dots, P$, it seems reasonable to consider a factorial model for the joint-probability density function of the observations as in

$$p_{\hat{\mathbf{X}}}(\mathbf{x}) = p_{\mathcal{S}_{1:P}}(\mathbf{y}) f(\mathbf{y}_c|\mathbf{y}) \tag{37}$$

where $p_{\mathcal{S}_{1:P}}(\mathbf{y}) = p_{\mathcal{S}_1}(y_1) \cdots p_{\mathcal{S}_P}(y_P)$ is the known joint pdf of the subset of desired sources, while $f(\mathbf{y}_c|\mathbf{y})$ is a complementary pdf which condenses all the remaining ignorance about the model.

Assuming a stationary i.i.d. vector process of observations $\{\mathbf{X}(t)\}$ and a sufficiently long sequence of samples drawn from it, by the weak law of large numbers, the normalized log likelihood converges in probability to

$$L(\mathbf{U}, f|p_{\mathcal{S}_{1:P}}) = \int p_{\mathbf{X}}(\mathbf{x}) \log p_{\hat{\mathbf{X}}}(\mathbf{x}) d\mathbf{x}. \tag{38}$$

By maximizing this function, one automatically minimizes the divergence between the true pdf of the observations and that of the estimate, as a consequence of the following relationship between both quantities

$$L(\mathbf{U}, f|p_{\mathcal{S}_{1:P}}) = -D(p_{\mathbf{X}} \| p_{\hat{\mathbf{X}}}) - h(\mathbf{X}). \tag{39}$$

The maximum likelihood with respect to the unknown pdf $f(\cdot)$ is obtained by decomposing the divergence

$$L(\mathbf{U}, f|p_{\mathcal{S}_{1:P}}) = -D(p_{\mathbf{Y}} \| p_{\mathcal{S}_{1:P}}) - D\left(p_{\mathbf{Y}_c|\mathbf{Y}} \| f\right) - h(\mathbf{X}) \tag{40}$$

and noting that $D(p_{\mathbf{Y}_c|\mathbf{Y}} \| f) \ge 0$, with equality only if $f^* = p_{\mathbf{Y}_c|\mathbf{Y}}$ almost everywhere. Thus, the best estimate with respect to $f(\cdot)$ is

$$p_{\hat{\mathbf{X}}}^{(\text{ML})}(\mathbf{x}) = p_{\mathcal{S}_{1:P}}(\mathbf{y}) p_{\mathbf{Y}_c|\mathbf{Y}}(\mathbf{y}_c|\mathbf{y}). \tag{41}$$

Substituting this result in the log-likelihood function and removing the constant term gives the maximum-likelihood contrast function for extraction of the desired sources

$$\begin{aligned} Q_{\text{ML}}(\mathbf{Y}|p_{\mathcal{S}_{1:P}}) &= L(\mathbf{U}, f^*|p_{\mathcal{S}_{1:P}}) + h(\mathbf{X}) \\ &= -D(p_{\mathbf{Y}} \| p_{\mathcal{S}_{1:P}}) \end{aligned} \tag{42}$$

subject to $\text{Cov}(\mathbf{Y}) = \text{Cov}(\mathcal{S}_{1:P}) = \mathbf{I}_P$. Thus, maximizing the likelihood of the observations is equivalent to minimizing the Kullback–Leibler divergence between the joint density of the outputs and that of the desired sources. Rewriting the maximum-likelihood contrast function as

$$Q_{\text{ML}}(\mathbf{Y}|p_{\mathcal{S}_{1:P}}) = h(\mathbf{Y}) + \sum_{i=1}^{P} \mathbf{E}[\log p_{\mathcal{S}_i}(y_i)] \tag{43}$$

we note that it depends not only on the densities of the sources, but also on the joint differential entropy of the outputs $h(\mathbf{Y})$, whose optimization is now much more difficult than in the BSS case and, for $P > 1$, involves working with multivariate statistics. However, considering an upper bound for the ML contrast function, one obtains the *marginal maximum-likelihood* (MML) contrast function

$$Q_{\text{MML}}(\mathbf{Y}|p_{\mathcal{S}_{1:P}}) = Q_{\text{ML}}(\mathbf{Y}|p_{\mathcal{S}_{1:P}}) + I(Y_1, \dots, Y_P) \tag{44}$$

$$= -\sum_{i=1}^{P} D(p_{Y_i} \| p_{\mathcal{S}_i}) \tag{45}$$

for which only univariate marginal densities are necessary. The global maximum of $Q_{\text{MML}}(\mathbf{Y}|p_{\mathcal{S}_{1:P}})$ is only attained when we extract the sources with the desired densities. The proof of this result follows from the application of the Darmois–Skitovitch theorem [16] together with the observation that each term $-D(p_{Y_i} \| p_{\mathcal{S}_i})$ is always nonpositive and it reaches the maximum value (which is zero) only if $p_{Y_i} = p_{\mathcal{S}_i}$ almost everywhere.

Fig. 2 graphically illustrates the relationship between the different criteria for BSE. The divergences between densities in the plot play the role of squared distances [29], [31]. In the figure, $p_{\mathcal{S}_{1:P}}$ refers to the joint density of the P sources with the lowest differential entropy. Thus, the position of the joint density of the observations $p_{\mathbf{Y}}(\mathbf{y})$ is upper bounded by a hypersphere with its center in the Gaussian density and squared radius $\Psi_{\text{Neg}}(\mathcal{S}_{1:P})$. If the P observations are a linear combination of only P sources,

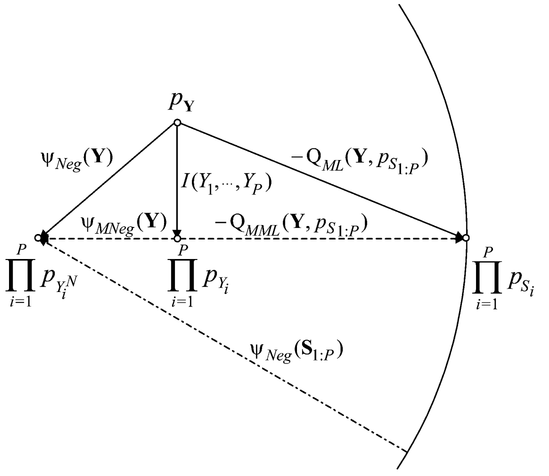
$p_{\mathbf{Y}}$
$\Psi_{Neg}(\mathbf{Y})$
$I(Y_1,\cdots,Y_P)$
$-\mathrm{Q}_{ML}(\mathbf{Y},p_{S_{1:P}})$
$\Psi_{MNeg}(\mathbf{Y})$
$-\mathrm{Q}_{MML}(\mathbf{Y},p_{S_{1:P}})$
$\prod_{i=1}^{P} p_{Y_i^N}$
$\prod_{i=1}^{P} p_{Y_i}$
$\prod_{i=1}^{P} p_{S_i}$
$\Psi_{Neg}(\mathbf{S}_{1:P})$

Fig. 2. The Pythagorean decomposition of the divergence illustrates the relationship between the different criteria. The drawing illustrates the case in which $N > P > 1$. The divergences between densities in the plot play the role of squared distances. The asymmetry of the Kullback–Leibler divergence $D(\cdot||\cdot)$ is reflected using arrows which depart from the first density toward the second density.

we have a reduced BSS problem for which $p_{\mathbf{Y}}(\mathbf{y})$ is located at some point on the hypersphere. It is clear from the figure that the maximization of one of the following criteria: $\Psi_{M\mathrm{Neg}}(\mathbf{Y})$, $Q_{\mathrm{ML}}(\mathbf{Y}, p_{\mathcal{S}_{1:P}})$ and $Q_{\mathrm{MML}}(\mathbf{Y}, p_{\mathcal{S}_{1:P}})$, solves the BSE problem ($p_{\mathbf{Y}}$ converges to $p_{\mathcal{S}_{1:P}}$); whereas the maximization of $\Psi_{\mathrm{Neg}}(\mathbf{Y})$ or the minimization of $I(Y_1, \ldots, Y_P)$ does not.

B. Information Maximization

For BSS, the connection between the information maximization principle and maximum likelihood criterion is supported in [21]. A similar connection still holds true for BSE, but with an important difference with respect to the BSS case; the blind form of the infomax criteria does not give a valid contrast function for BSE.

The information maximization principle (infomax) was proposed by Linsker [17] and was motivated by analysis of the sensory system. The principle suggests that the layers of a sensory network are adapted to the environment in their attempt to preserve as much information as possible, which is achieved by maximizing the transfer of information through the layer of neurons. The extraction system followed by a bank of nonlinearities $T_i(Y_i)$, $i = 1, \ldots, P$, is a layer of P neurons.

The infomax principle consists in maximizing the differential entropy of the nonlinear, bounded transformation $T(\mathbf{Y}) = [T_1(Y_1), \cdots, T_P(Y_P)]^T$ of the outputs, i.e., maximizing

$$Q_{\mathrm{Inf}}(\mathbf{Y}|T) = h(T(\mathbf{Y})) \tag{46}$$

Nadal and Parga proved [18] that the maximum of the infomax principle, with respect to the functional form of the nonlinearity, is obtained when each $T_i(\cdot)$ matches with $T_i^*(\cdot)$, the cumulative distribution function of the corresponding output Y_i. This result can be better understood after rewriting the *infomax* index as the opposite of the divergence of the joint density $p_{T(\mathbf{Y})}$ from the P-dimensional uniform density $p_{\mathcal{U}_{1:P}}$ with support in $[0,1] \times \ldots \times [0,1]$ (note that $h(\mathcal{U}_{1:P})$ is zero) and using the Pythagorean decomposition of this divergence

$$\begin{aligned} Q_{\mathrm{Inf}}(\mathbf{Y}|T) &= -D\left(p_{T(\mathbf{Y})}||p_{\mathcal{U}_{1:P}}\right) \qquad (47)\\ &= -I(Y_1, \ldots, Y_P)\\ &\quad -\sum_{i=1}^{P} D\left(p_{T_i(Y_i)}||p_{T_i^*(Y_i)}\right) \qquad (48) \end{aligned}$$

The first term on the right-hand side accounts for the lack of independence of the outputs, whereas the second term accounts for the departure from the optimal nonlinearities. Based on these ideas, Bell and Sejnowski developed a successful implementation of the *infomax algorithm* for ICA [19]. In a BSE case, even for the optimal nonlinearity (i.e., when $Q_{\mathrm{Inf}}(\mathbf{Y}|T^*) = -I(Y_1, \ldots, Y_P)$), the independence of the outputs does not guarantee the extraction of any of the sources (see Section III), and the infomax principle fails. However, when one uses the *a priori* information about the desired densities to constrain the form of the nonlinearity, the infomax approach reduces to the ML contrast function suitable for the extraction of the sources. Let $T_{\mathcal{S}_{1:P}}(\cdot)$ denote the joint-cumulative distribution function of the P independent sources we want to extract, then it holds

$$\begin{aligned} Q_{\mathrm{Inf}}\left(\mathbf{Y}|T_{\mathcal{S}_{1:P}}\right) &= -I(Y_1, \ldots, Y_P)\\ &\quad -\sum_{i=1}^{P} D\left(p_{T_{S_i}(Y_i)}||p_{U_i}\right)\\ &= Q_{\mathrm{ML}}\left(\mathbf{Y}|p_{\mathcal{S}_{1:P}}\right). \end{aligned} \tag{49}$$

Adding $I(Y_1, \ldots, Y_P)$ one obtains the marginal form of the contrast function (which only involves univariate statistics)

$$\begin{aligned} Q_{M\mathrm{Inf}}\left(\mathbf{Y}|T_{\mathcal{S}_{1:P}}\right) &= -\sum_{i=1}^{P} D\left(p_{T_{S_i}(Y_i)}||p_{U_i}\right)\\ &= Q_{\mathrm{MML}}\left(\mathbf{Y}|p_{\mathcal{S}_{1:P}}\right). \end{aligned} \tag{50}$$

There is a striking parallel in the relationship between *infomax* and the *maximum likelihood* criteria, and that between the divergence from the uniform distribution (*minimum information*) and the *negentropy* entropy criteria. But, at the same time, there is a clear difference, since *infomax* maximizes the differential entropy of the transformed random variable (46), whereas the divergence from the uniform distribution (24) minimizes it.

C. The Entropy-Likelihood Criterion

By analyzing the ML-density estimate $p_{\hat{\mathbf{X}}}^{(\mathrm{ML})}(\mathbf{x})$ of (41), one can see that it uses the conditional density $p_{\mathbf{Y}_c|\mathbf{Y}}$, which is unavailable information. Therefore, this density should be replaced by the least biased density estimate which is consistent with the normalized mean and covariance, i.e., the maximum entropy estimate $p_{\mathbf{Y}_c^{\mathcal{N}}|\mathbf{Y}^{\mathcal{N}}}$. With this substitution, one obtains the entropy-likelihood estimate

$$p_{\hat{\mathbf{X}}}^{(\mathrm{EL})}(\mathbf{x}) = p_{\mathcal{S}_{1:P}}(\mathbf{y})p_{\mathbf{Y}_c^{\mathcal{N}}|\mathbf{Y}^{\mathcal{N}}}(\mathbf{y}_c|\mathbf{y}). \tag{51}$$

In similarity with the PPDE approach, we consider as a contrast function the relative improvement in the fit to

TABLE I
ESTIMATES OF THE JOINT DENSITY OF THE OBSERVATIONS, ASSOCIATED WITH THE DIFFERENT CRITERIA FOR THE EXTRACTION OF P SOURCES

CRITERIA	$p_{\hat{\mathbf{X}}}(\mathbf{x})$
Marginal negentropy	$\left(\prod_{i=1}^{P} p_{Y_i}(y_i)\right) p_{\mathbf{Y}_c^{\mathcal{N}}\lvert\mathbf{Y}^{\mathcal{N}}}(\mathbf{y}_c\lvert\mathbf{y})$
Entropy-likelihood	$\left(\prod_{i=1}^{P} p_{\mathcal{S}_i}(y_i)\right) p_{\mathbf{Y}_c^{\mathcal{N}}\lvert\mathbf{Y}^{\mathcal{N}}}(\mathbf{y}_c\lvert\mathbf{y})$
ML & Infomax	$\left(\prod_{i=1}^{P} p_{\mathcal{S}_i}(y_i)\right) p_{\mathbf{Y}_c\lvert\mathbf{Y}}(\mathbf{y}_c\lvert\mathbf{y})$

the true density of the observations, between the initial estimate $p_{\hat{\mathbf{X}}}^{(0)}(\mathbf{x}) = p_{\mathbf{X}^{\mathcal{N}}}(\mathbf{x})$ and the improved estimate $p_{\hat{\mathbf{X}}}^{(1)}(\mathbf{x}) = p_{\hat{\mathbf{X}}}^{(\mathrm{EL})}(\mathbf{x})$ (which results from the knowledge of the densities of the desired sources)

$$Q_{\mathrm{EL}}\left(\mathbf{Y}|p_{\mathcal{S}_{1:P}}\right) = D\left(p_{\mathbf{X}}\|p_{\hat{\mathbf{X}}}^{(0)}\right) - D\left(p_{\mathbf{X}}\|p_{\hat{\mathbf{X}}}^{(1)}\right) \quad (52)$$

$$= D\left(p_{\mathbf{Y}}\|p_{\mathbf{Y}^{\mathcal{N}}}\right) - D\left(p_{\mathbf{Y}}\|p_{\mathcal{S}_{1:P}}\right) \quad (53)$$

$$= \frac{P}{2}\log(2\pi e) + \sum_{i=1}^{P} \mathbf{E}\left[\log p_{\mathcal{S}_i}(y_i)\right]. \quad (54)$$

This function, which resembles the contrast function for extraction proposed in [39] from a different perspective, is very attractive because the density estimation of the outputs is no longer needed. A novel interpretation of the contrast function results from the observation that its associated criterion is a combination of the *negentropy* and *maximum likelihood* criteria,

$$Q_{\mathrm{EL}}\left(\mathbf{Y}|p_{\mathcal{S}_{1:P}}\right) = \psi_{\mathrm{Neg}}(\mathbf{Y}) + Q_{\mathrm{ML}}\left(\mathbf{Y}|p_{\mathcal{S}_{1:P}}\right). \quad (55)$$

Hence, we name this criterion the *entropy-likelihood*. In (55) the first term tries to increase the divergence from Gaussianity, whereas the second one tries to fit the distributions of the outputs to the desired ones. Both objectives are compatible (simultaneously maximized) if the desired sources are those with the lowest differential entropy in the mixture, i.e., $h(S_i) \leq h(S_j)$ for $1 \leq i \leq P < j \leq N$, and, in this case, the extraction solution is a global maximum of $Q_{\mathrm{EL}}(Y_1|p_{\mathcal{S}_i})$. When this condition is not satisfied, the entropy-likelihood contrast function still holds locally in the vicinity of the desired sources (see the discussion after corollary 1 in Section VIII), but the desired extraction solution is only guaranteed to be a local maximum of $Q_{\mathrm{EL}}(\mathbf{Y}|p_{\mathcal{S}_{1:P}})$.

Table I summarizes the different estimates of the joint pdf of the observations, associated with the criteria for the extraction of P sources.

VII. THE EXTRACTION ALGORITHM AND THE NON-LINEAR (SCORE) FUNCTIONS

A particularly simple and useful method to maximize any chosen contrast function subject to the constraint $\mathrm{Cov}(\mathbf{Y}) = \mathbf{I}_P$ is to use the natural gradient ascent in the Stiefel manifold of semiorthogonal matrices [38], which is given by

$$\tilde{\nabla}_{\mathbf{U}}\Psi = \nabla_{\mathbf{U}}\Psi - \mathbf{U}(\nabla_{\mathbf{U}}\Psi)^T\mathbf{U} \quad (56)$$

where $\nabla_{\mathbf{U}}\Psi$ is the usual gradient. The application of the natural gradient algorithm to solve the problem of simultaneous BSE was proposed by Amari [39]. Using the chain rule, one can see that

$$\nabla_{\mathbf{U}}\Psi = \left[\frac{\partial\Psi(\mathbf{Y})}{\partial U_{ij}}\right]_{ij} = -\mathbf{D}\mathbf{R}_{\varphi,x} \quad (57)$$

where $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_P)$ is a diagonal matrix of ordering constants, $\mathbf{R}_{\varphi,x} = \mathrm{E}_t[\varphi(\mathbf{y})\mathbf{x}^T]$ is the sample cross-correlation matrix between specific nonlinearities and the observations, and the vector of nonlinearities $\varphi(\mathbf{y}) = [-(d\tilde{\psi}_1(y_1)/dy_1), \ldots, -(d\tilde{\psi}_P(y_P)/dy_P)]^T$ is a vector function which depends on $\tilde{\psi}_i(\cdot)$ (the stochastic form of the indexes $\psi_i(Y_i) = \mathrm{E}[\tilde{\psi}_i(Y_i)]$).

The resulting natural gradient algorithm takes the following simple form:

$$\mathbf{U}^{(n+1)} = \mathbf{U}^{(n)} + \mu\tilde{\nabla}_{\mathbf{U}}\Psi \quad (58)$$

$$= \mathbf{U}^{(n)} - \mu\left(\mathbf{D}\mathbf{R}_{\varphi,x}^{(n)} - \mathbf{R}_{y,\varphi}^{(n)}\mathbf{D}\mathbf{U}^{(n)}\right). \quad (59)$$

The nonlinearities for the *entropy-likelihood* contrast function are the score functions of the densities of the desired sources $\varphi_i^{(\mathrm{EL})}(y_i) = -(p'_{S_i}(y_i)/p_{S_i}(y_i))$, so they can be computed when this information is known. This is not necessary for the *marginal-negentropy* contrast function since approximations for the nonlinearities $\varphi_i^{(M\mathrm{Neg})}(y_i) = -(p'_{Y_i}(y_i)/p_{Y_i}(y_i))$ have been obtained in [15] and in [22] by using the truncated Edgeword and Gram–Charlier expansions of the marginal pdfs of the outputs in the vicinity of the Gaussian distribution (a different estimation procedure is presented in the appendix of [4]). The nonlinearities for the *marginal maximum-likelihood* criteria take the form of $\varphi_i^{(\mathrm{MML})}(y_i) = \varphi_i^{(\mathrm{EL})}(y_i) - \varphi_i^{(M\mathrm{Neg})}(y_i)$. However, in practice, these estimates may not always perform well, since close to the extraction of any of the sources the distribution of the output is far from being close to the Gaussian and the truncated expansions no longer result accurate.

A good alternative approach is the cumulant-based index, because for it, the general form of the nonlinearity can be obtained without approximations, and it is universal in the sense that it will work under the weak condition that each of the desired independent components has a nonzero index. The nonlinearity of the cumulant index is a linear combination of partial nonlinearities $\varphi^{(r)}(y_i)$, where each one is related to the rth-order cumulant, i.e.,

$$\varphi(y_i) = \sum_{r>2}\omega_r\varphi^{(r)}(y_i) \quad (60)$$

The expressions of the partial nonlinearities are explicitly shown in Table II up to order seven, although, in practice, cumulants with order $r \geq 5$ are not usually used by themselves, but as complementary information, since their precise estimation requires a large number of samples.

Our objective is to extract the desired independent components, i.e., the source signals with the largest indexes $\psi_i(S_i)$. Since we use a gradient algorithm, it can be trapped in the local maxima of the contrast function corresponding to other valid extracting solutions or to defective solutions (provided they exist). Therefore, in general, there is no guarantee that one will always achieve the global maximum solution in one single stage of extraction, and thus, the local search should be combined with

TABLE II
PARTIAL ACTIVATION FUNCTIONS $\varphi^{(r)}(\cdot)$ ASSOCIATED WITH rth-ORDER CUMULANTS

$\varphi^{(3)}(y_i) = -\mathtt{sign}(C^3_{y_i})\ \lvert C^3_{y_i}\rvert^{(\alpha_3-1)}\ y_i^2$
$\varphi^{(4)}(y_i) = -\mathtt{sign}(C^4_{y_i})\ \lvert C^4_{y_i}\rvert^{(\alpha_4-1)}\ (y_i^3 - 3y_i\mathtt{E}[y_i^2])$
$\varphi^{(5)}(y_i) = -\mathtt{sign}(C^5_{y_i})\ \lvert C^5_{y_i}\rvert^{(\alpha_5-1)}\ (y_i^4 - 4y_i\mathtt{E}[y_i^3] - 6y_i^2\mathtt{E}[y_i^2])$
$\varphi^{(6)}(y_i) = -\mathtt{sign}(C^6_{y_i})\ \lvert C^6_{y_i}\rvert^{(\alpha_6-1)}\ (y_i^5 - 5y_i\mathtt{E}[y_i^4] - 10y_i^2\mathtt{E}[y_i^3] - 10y_i^3\mathtt{E}[y_i^2] + 30y_i(\mathtt{E}[y_i^2])^2)$
$\varphi^{(7)}(y_i) = -\mathtt{sign}(C^7_{y_i})\ \lvert C^7_{y_i}\rvert^{(\alpha_r-1)}\ (y_i^6 - 6y_i\mathtt{E}[y_i^5] - 15y_i^2\mathtt{E}[y_i^4] - 20y_i^3\mathtt{E}[y_i^3] - 15y_i^4E[y_i^2] + 120y_i\mathtt{E}[y_i^2]\mathtt{E}[y_i^3] + 90y_i^2(\mathtt{E}[y_i^2])^2)$

some kind of global search procedure and test for the validity of the solution. When the desired sources are extracted, one can stop the search. Otherwise, the extraction procedure can be repeated starting from a different initial condition or after performing the deflation of the extracted components (see [33] for more details on deflation). The procedure can be stopped when one recovers the desired sources or when all the recovered independent components in the last extractions exhibit small indexes.

VIII. STABILITY ANALYSIS OF THE ALGORITHM

In this section, we study the local convergence of the algorithm. We consider an arbitrary vector of nonlinearities $\varphi(\mathbf{Y}) = [\varphi_1(Y_1), \ldots, \varphi_P(Y_P)]$ and denote the ith nonlinearity briefly as $\varphi_i = \varphi_i(S_i)$ when acting on the corresponding extracted source.

When a sufficiently small step size μ is used, the extraction solutions should be attractors for the gradient algorithm, provided that they are maxima of the corresponding semiorthogonal contrast function. However, the imprecise estimation of the nonlinearity φ_i associated with the function $\psi_i(.)$ sometimes changes the status of the contrast function in such a way that the approximated function $\hat{\psi}_i(.)$ no longer produces a maximum in the extraction solution, but another kind of critical point. This is one of the reasons that justifies interest in the analysis of the local convergence of the algorithm for an arbitrary nonlinearity. A second reason is that it is useful to establish possible adequate step sizes that ensure a high convergence rate and simultaneously guarantee the stability of the algorithm. The next theorem presents bounds for the learning step size resulting from the asymptotical stability analysis of the algorithm.

Theorem 2: Assuming that the mixing system is orthogonal, the necessary and sufficient local stability conditions of the gradient algorithm in the Stiefel manifold (59) to converge to a true solution are, for all $i, j|_{i\neq j} = 1, \ldots, P$, given by

$$0 < \mu < \frac{2}{d_i\kappa_i} \qquad \text{if} \quad P = 1 \tag{61}$$

$$0 < \mu < \min\left\{\frac{2}{d_i\kappa_i + d_j\kappa_j}, \frac{2}{d_i\kappa_i}\right\} \quad \text{if1} < P < N \tag{62}$$

$$0 < \mu < \frac{2}{d_i\kappa_i + d_j\kappa_j} \qquad \text{if} \quad P = N \tag{63}$$

where the variables $\kappa_i = \kappa_i(S_i)$ that control the local stability are given by[1]

$$\kappa_i = \mathbf{E}\left[\frac{\partial\varphi_i(S_i)}{\partial s_i}\right]\mathrm{Cov}(S_i) - \mathbf{E}\left[S_i\varphi_i(S_i)\right]. \tag{64}$$

The proof of the theorem can be found in the Appendix and is based on the analysis of the linearized dynamic of the algorithm around the extraction solution. A similar study was previously used in [40] to find the local stability conditions of the EASI algorithm. In our analysis we focus on the BSE algorithm, and we provide bounds for the learning step size. The linearized analysis of the algorithm also reveals a simple estimate for the step size that guarantees stability and a fast convergence. If $\kappa_i > 0\ \forall i = 1, \ldots, P$, a good (close to the optimum) candidate for the step size μ at iteration n is

$$\mu^{(n)} = \frac{1}{2\max_i|\kappa(Y_i)|}. \tag{65}$$

The results of the following corollaries, when substituted in (65) and in (61)–(63), assist in clarifying the choice and the bounds for the learning step size.

Corollary 1: When the nonlinear functions match the score functions of the distribution of the extracted sources $\varphi_i = -(\partial \log p_{S_i}(s_i)/\partial s_i)$ in a local neighborhood of the extraction solution, the factors that control the local stability of the algorithm can be expressed as

$$\kappa_i \overset{(a)}{=} \mathrm{Cov}(S_i)E\left[\left(\frac{\partial \log p_{S_i}(s_i)}{\partial s_i}\right)^2\right] - 1 \tag{66}$$

$$\overset{(b)}{=} \mathrm{Cov}(S_i)E\left[\left(\frac{\partial}{\partial s_i}\log\frac{p_{S_i}(s_i)}{p_{S_i^{\mathcal{N}}}(s_i)}\right)^2\right] \tag{67}$$

$$\geq 0. \tag{68}$$

The proof of this corollary is presented in the Appendix and its interpretation is the following. In (a) the definition of κ_i takes the form of standardized Fisher information, whereas in (b) its interpretation is stressed as a factor that expresses deviation from Gaussianity (note that $\kappa_i = 0$ only for the Gaussian distribution). In fact, this function $\kappa_i(\cdot)$ is by itself a contrast function maximized by one of the independent components in the mixture and is minimized by the Gaussian distribution [10]. From the corollary we observe that, for the proper step sizes, the gradient ascent algorithm is locally convergent to any of

[1] The κ_i factors were originally defined in [40]. Their definition in Theorem 2 includes the covariance of the sources only for theoretical purposes, since we have adopted throughout the paper the normalization $\mathrm{Cov}(S_i) = 1\ \forall i$.

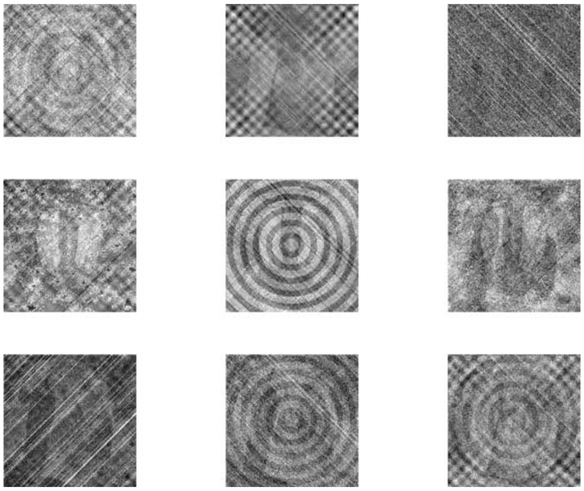

Fig. 3. Images of the nine observations after prewhitening.

the extraction solutions consisting of non-Gaussian sources, evidencing that all these solutions are local maxima of the *marginal negentropy* and *entropy-likelihood* contrast functions.

Corollary 2: For the cumulant-based index the factors that control the local stability can be rewritten as

$$\kappa_i = \sum_{r>2} \omega_r \left| C^r_{S_i} \right|^{\alpha_r} \geq 0. \tag{69}$$

The proof of this corollary relies on the fact that the expectation of the derivative of the nonlinearity for the cumulant-based index is always zero ($\mathrm{E}[\partial\varphi_i/\partial s_i] = 0$) and, thus from the definition of κ_i, we obtain

$$\kappa_i = -\mathbf{E}[S_i\varphi_i] = \sum_{r>2} \omega_r \left| C^r_{S_i} \right|^{\alpha_r} \geq 0 \tag{70}$$

which is always nonnegative for all $i = 1, \ldots, P$, even when the densities of the sources are unknown. Again, the function $\kappa_i(\cdot)$ of the corollary is by itself a contrast function for blind extraction.

Incidentally, the behavior of the term $\mathrm{E}[\partial\varphi_i/\partial s_i]$ is substantially different in both corollaries (it is greater than the unity for corollary 1 and zero for corollary 2). This indicates that cumulant-based contrasts like (26), which do not involve cross-products of cumulants with different orders, cannot be used to construct accurate approximations of the *marginal negentropy* or of the *marginal maximum likelihood* contrasts.

IX. Simulations

In the first experiment, we illustrate an interesting theoretical behavior of the extraction algorithm. To facilitate graphical representation of the results, we consider the nine observed images ($N = 9$) shown in Fig. 3. These images are prewhitened versions of the original observations, so they satisfy the decorrelation constraint $\mathrm{Cov}(\mathbf{X}) = \mathbf{I}_N$. The observations were generated from a random linear combination of nine independent-source images, whose shapes are barely distinguishable in Fig. 3. The source images have different kurtosis signs, and two of them are very close to being Gaussian noise. The kurtosis of the sources $\{C^4_{s_i},\ i = 1, \ldots, 9\}$ are: {4.2, 3.5, −2, 1.9, −1.6, −1, −1, 0.05, 0.01}.

We chose the criteria based on higher order cumulants, i.e., (59) with $r = 4, \alpha_r = 1$. We set the number of sources to extract to $P = 3$, and applied a batch version of the natural gradient algorithm with the adaptive step size in (65). Using Corollary 2, one can see that the recommended step size takes the form

$$\mu^{(n)} = \frac{1}{2\max_i \left| C^4_{y_i} \right|}. \tag{71}$$

We started from an initialization $\mathbf{U}^{(0)} = [\mathbf{I}_P, \mathbf{0}_{P\times N-P}]$, which selects as initial outputs the P observed images of the first row. After 16 iterations, the algorithm converged to the first three extracted sources shown in Fig. 4(a). Then, if these are not the sources of interest we can remove the contribution of these sources from the observation and perform a new extraction. The second extraction started from the P observations of the next row, i.e., $\mathbf{U}^{(0)} = [\mathbf{0}_P, \mathbf{I}_P, \mathbf{0}_{P\times N-2P}]$, and convergence was obtained after 19 iterations to produce the three sources of Fig. 4(b). Finally, the third extraction started from the last P observations ($\mathbf{U}^{(0)} = [\mathbf{0}_{P\times N-P}, \mathbf{I}_P]$) and converged after 22 iterations to produce the three sources of Fig. 4(c). One can clearly see how the algorithm perfectly recovered the nine source images.

In agreement with the theoretical results, the algorithm attempts at the first stage to extract the most structured sources or less random in a certain sense (in this example those with the largest absolute value of kurtosis), whereas the sources that are closer to being Gaussian, or those with greater uncertainty, are typically extracted in the last stage of extractions. However, since the gradient algorithm has local scope, the result is also influenced by the initialization. For this reason, it is helpful to choose a good initialization condition that produces outputs as close as possible to the desired sources. This can be done directly by identifying those observations that provide better estimates of the desired sources and choosing them as initial outputs. This is what has been done in the previous experiment, one can compare Figs. 3 and 4 to see how initialization influences determination by the algorithm of which sources are recovered at each output.

Although the contrast function based on cumulants measures the departure of the outputs from Gaussianity, we still have some control to selectively extract source signals with specific stochastic properties through the proper selection of the involved cumulants orders r and the factors ω_r and α_r. For instance, if the sources of our interest have asymmetric distributions, we can favor their extraction in the first place by weighting more in the index (26) the skewness and other cumulants of odd order.

To illustrate this possibility, in a second simulation we consider random mixtures of 100 normalized sources. Only five of them are asymmetric binary sources with probability mass function $p_S(s) = 0.2\delta[s-2] + 0.8\delta[s+0.5]$, and the remaining 95 are binary symmetric sources with probability mass function $p_S(s) = 0.5\delta[s+1] + 0.5\delta[s-1]$. We favor the simultaneous extraction of the asymmetric sources from the mixture using 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 the number of sources to be extracted to $P = 5$ and performed 100 random simulations. The histogram was used to distinguish the desired sources

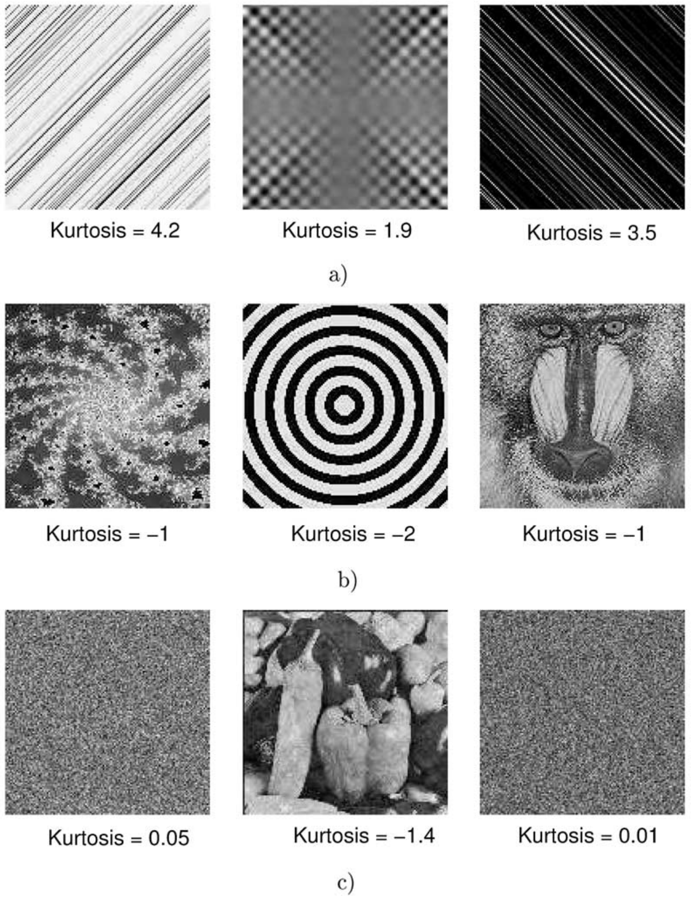

Fig. 4. Extraction of the nine sources in groups of three $(P = 3)$. After each extraction a deflation procedure has been applied. a) First extraction: 16 iterations, $\Psi_{\rm Cum}(\mathbf{y}) = 9.6$. b) Second extraction: 19 iterations, $\Psi_{\rm Cum}(\mathbf{y}) = 4$. c) Third extraction: 22 iterations, $\Psi_{\rm Cum}(\mathbf{y}) = 1.46$.

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. In 22% of the experiments 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 the 100% after the third run.

The interested reader can also apply these algorithms to his own data or compare them with other existing ones. The natural gradient algorithm for the contrast function based on cumulants has been implemented in the *ICALAB toolbox* [41] for MatLab under the algorithm name SIMBEC (Simultaneous BSE using Cumulants).

X. Conclusion

In this paper, we have presented a unified interpretation of several existing information theoretic criteria for BSE by using the projection pursuit density estimation methodology. Our main contribution is the development of some tools that allow the extension of these criteria (already known for extraction of a single source and for blind source separation) to the case of simultaneous blind extraction of an arbitrary number of sources $P \leq N$. The natural gradient algorithm in the Stiefel manifold is a suitable technique for the optimization of the semiorthogonal contrasts associated with these criteria. We have analyzed the local convergence of this algorithm and provided useful bounds for its learning step size. Finally, we have demonstrated with some sample experiments the validity of the theoretical results and the good performance of the proposed algorithm.

Appendix

A. Proof of Properties 1 and 2 for the Negentropy Contrast

If we express the normalized random output in terms of the global transfer system and the sources $Y_1 = \sum_{j=1}^{N} G_{1j}S_j$, we can apply the entropy power inequality (lemma 1) to see that

$$e^{2h(Y_1)} \geq \sum_{j=1}^{N} e^{2h(G_{1j}S_j)} = \sum_{j=1}^{N} |G_{1j}|^2 e^{2h(S_j)}. \tag{72}$$

After taking logarithms, we can use this result to upper bound $\psi_{\rm Neg}(Y_1)$ as

$$\begin{aligned}\psi_{\rm Neg}(Y_1) &= \frac{1}{2}\log(2\pi e) - h(Y_1)\\ &\overset{(a)}{\leq} \frac{1}{2}\log(2\pi e) - \frac{1}{2}\log\left(\sum_{j=1}^{N}|G_{1j}|^2 e^{2h(S_j)}\right)\\ &\overset{(b)}{\leq} \sum_{j=1}^{N}|G_{1j}|^2\left(\frac{1}{2}\log(2\pi e) - h(S_j)\right)\end{aligned} \tag{73}$$

where the inequality (a) originates from (72) whereas the inequality (b) is a consequence of the strict concavity of the logarithm and of the normalization of the global transfer system $\sum_{j=1}^{N} |G_{1j}|^2 = 1$. Therefore, we have proved the desired property

$$\psi_{\text{Neg}}(Y_1) \le \sum_{j=1}^{N} |G_{1j}|^2 \psi_{\text{Neg}}(S_j) \tag{74}$$

with equality only when Y_1 is one of the independent sources (being for this case $\psi_{\text{Neg}}(Y_1)$ positive) or is an arbitrary linear combination of Gaussian sources (being for this case, $\psi_{\text{Neg}}(Y_1) = 0$).

The proof of property 2 stems directly from the properties of the Kullback–Leibler divergence, which is always nonnegative and equal to zero, only if the two arguments (distributions) match almost everywhere.

B. Proof of Properties 1 and 2 for the Cumulant Based Contrast Functions

The proof of property 1 is obtained by bounding the rth higher order cumulant $(r > 2)$ of the output by

$$\left|C^r_{Y_1}\right|^{\alpha_r} \overset{(a)}{=} \left|\sum_{j=1}^{N} G^r_{1j} C^r_{S_j}\right|^{\alpha_r} \tag{75}$$

$$\overset{(b)}{\le} \left|\sum_{j=1}^{N} |G_{1j}|^2 \left|C^r_{S_j}\right|\right|^{\alpha_r} \tag{76}$$

$$\overset{(c)}{\le} \sum_{j=1}^{N} |G_{1j}|^2 \left|C^r_{S_j}\right|^{\alpha_r} \tag{77}$$

where the equality (a) follows from the properties of the cumulants [37]. The inequality (b) results from the fact that the global system is of unit norm and therefore $|G_{1j}|^r \le |G_{1j}|^2 \le 1$. The last inequality (c) holds true trivially for $\alpha_r = 1$ and follows from the convexity of the power function $|\cdot|^{\alpha_r}$ for $\alpha_r > 1$.

Property 2 states that the contrast function should be zero for the Gaussian distribution, i.e., $\psi_{\text{Cum}}(Y_1^{\mathcal{N}}) = 0$. This is easily verified since the higher order cumulants of Gaussian processes are all zero.

C. Proof of Theorem 1

The decorrelation constraint for the outputs $(\text{Cov}(\mathbf{Y}) = \mathbf{G}\mathbf{G}^T = \mathbf{I}_P)$ is tantamount to the semiorthogonality of the global transfer matrix $\mathbf{G}$. Let us define the diagonal matrices $\mathbf{D} = \text{diag}(d_1, \ldots, d_P)$ and $\mathbf{\Lambda} = \text{diag}(\psi(S_1), \ldots, \psi(S_N))$. From property 2 we have that

$$\sum_{i=1}^{P} d_i \psi(Y_i) \le \sum_{i=1}^{P} d_i \sum_{j=1}^{N} |G_{ij}|^2 \psi(S_j) = \text{trace}\{\mathbf{DM}\} \tag{78}$$

where $\mathbf{M} = \mathbf{G\Lambda G}^T$ is a symmetric matrix of eigenvalues $\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_P$ and whose ordered diagonal elements are $m_1 \ge m_2 \ge \ldots \ge m_P$.

The proof of the theorem consists of the following main steps:

$$\sum_{i=1}^{P} d_i \psi(Y_i) \overset{(a)}{\le} \sum_{i=1}^{P} d_i m_i \tag{79}$$

$$\overset{(b)}{\le} \sum_{i=1}^{P} d_i \sigma_i \tag{80}$$

$$\overset{(c)}{\le} \sum_{i=1}^{P} d_i \psi(S_i). \tag{81}$$

The first inequality (a) is just a consequence of (78), since $\text{trace}\{\mathbf{DM}\} = \sum_{i=1}^{P} d_i m_i$. To prove the second inequality (b) we resort to the fact that $(m_1, \ldots, m_P) \prec (\sigma_1, \ldots, \sigma_P)$, i.e., the ordered set of diagonal elements of $\mathbf{M}$ are majorized by its eigenvalues, which means that for $k = 1, \ldots, P-1$

$$\sum_{i=1}^{k} (\sigma_i - m_i) \ge 0 \quad \text{and} \quad \sum_{i=1}^{P} (\sigma_i - m_i) = 0. \tag{82}$$

Thus, taking into account this majorization property and ordering of the constants $d_1 \ge d_2 \ge \cdots \ge d_P > d_{P+1} = 0$, we prove (b) since

$$\sum_{k=1}^{P} d_k(\sigma_k - m_k) = \sum_{k=1}^{P} (d_k - d_{k+1}) \sum_{i=1}^{k} (\sigma_i - m_i) \ge 0$$

where the last inequality follows from (82) and the fact that $(d_k - d_{k+1}) \ge 0$ for all $k = 1, \ldots, P$.

To prove the inequality (c) (81), we must resort to Poincaré's separation theorem of matrix algebra, which states that for a symmetric matrix $\mathbf{M} = \mathbf{G\Lambda G}^T$, with eigenvalues $\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_P$ under the constraint $\mathbf{GG}^T = \mathbf{I}_P$, the diagonal elements $\psi(S_1) \ge \psi(S_2) \ge \ldots \ge \psi(S_N)$ of $\mathbf{\Lambda}$ bound the eigenvalues of $\mathbf{M}$, satisfying $\forall i = 1, \ldots, P$

$$\psi(S_{N-P+i}) \le \sigma_i \le \psi(S_i). \tag{83}$$

Therefore the maximum of (78), subject to the semiorthogonality of $\mathbf{G}$, is

$$\max_{\mathbf{GG}^T = \mathbf{I}_P} \sum_{i=1}^{P} d_i \sigma_i = \sum_{i=1}^{P} d_i \psi(S_i). \tag{84}$$

If $\psi(S_1) > \ldots > \psi(S_P)$, the maximum is only obtained for those matrices $\mathbf{G}$ whose rows consist of orthogonal vectors that span the same subspace of the rows of $[\mathbf{I}_P, \mathbf{0}]$, which enforces $G_{ij} = 0\ \forall j > P$. From the weak form of strict convexity that satisfies $\psi(\cdot)$, which applies to the case of $\psi(\cdot) \neq 0$, the necessary and sufficient condition for the equality between (79) and (81) 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 $\psi(S_1) \ge \ldots \ge \psi(S_P)$, with equality for certain subsets of the first P sources which have a common value or index under $\psi(\cdot)$, the necessary and sufficient condition for the equality between (79) and (81) 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.

D. Linearized Dynamic of the Extraction Algorithm

Rewriting the natural gradient algorithm of (59) in terms of the global transfer system $\mathbf{G} = \mathbf{U}\mathbf{A}$ one finds

$$\begin{aligned}\mathbf{G}^{(n+1)} &= \mathbf{G}^{(n)} - \mu\Delta_{\mathbf{G}}^{(n)} \\ &= \mathbf{G}^{(n)} - \mu\mathbf{G}^{(n)} \\ &\quad \times \left(\left(\mathbf{R}_{s,\varphi}^{(n)}\mathbf{D}\mathbf{G}^{(n)}\right)^T - \mathbf{R}_{s,\varphi}^{(n)}\mathbf{D}\mathbf{G}^{(n)}\right). \end{aligned} \tag{85}$$

Since $\mathbf{G}^{(n)}\Delta_{\mathbf{G}}^{T(n)} + \Delta_{\mathbf{G}}^{(n)}\mathbf{G}^{T(n)} = 0$, the algorithm preserves the first order semiorthogonality of the global transfer system $\mathbf{G}^{(n+1)}\mathbf{G}^{T(n+1)} = \mathbf{G}^{(n)}\mathbf{G}^{T(n)} + o(\|\Delta_{\mathbf{G}}^{(n)}\|)$.

In the vicinity of the extraction solution ($\mathbf{G}_* = [\mathbf{I}_E, \mathbf{0}]$), the linearized form of the iteration completely determines the local stability behavior of the algorithm. Let us consider that the vector of N independent sources present in the mixture is split into two parts $\mathbf{S} = [\mathbf{S}_L^T, \mathbf{S}_R^T]^T$, where $\mathbf{S}_L = [S_1, \ldots, S_P]^T$ denotes the random vector of sources that are extracted by the algorithm, whereas, the vector $\mathbf{S}_R = [S_{P+1}, \ldots, S_N]^T$ contains the remaining ones.

Let us perturb the extraction solution by an additive matrix $\epsilon^{(n)} = [\epsilon_L^{(n)}, \epsilon_R^{(n)}]$, of arbitrary small norm, which preserves the first-order orthogonality of the global system, i.e., $\mathbf{G}^{(n)}\mathbf{G}^{T(n)} = \mathbf{I}_E + o(\|\epsilon^{(n)}\|)$, where the perturbation $\epsilon_L^{(n)}$ is skew-symmetric up to the first order, satisfying $\epsilon_L^{(n)} + \epsilon_L^{T(n)} = o(\|\epsilon^{(n)}\|)$. The resulting outputs are given by

$$\mathbf{Y} = \mathbf{G}^{(n)}\mathbf{S} = \mathbf{G}_L^{(n)}\mathbf{S}_L + \mathbf{G}_L^{(n)}\mathbf{S}_R \tag{86}$$

where $\mathbf{G}^{(n)} = \mathbf{G}_* + \epsilon^{(n)}$, $\mathbf{G}_L^{(n)} = \mathbf{I}_E + \epsilon_R^{(n)}$ and $\mathbf{G}_R^{(n)} = \epsilon_R^{(n)}$.

The first order Taylor expansion of function $\varphi(\mathbf{Y})$ at the extraction is

$$\varphi(\mathbf{Y}) = \varphi(\mathbf{S}_L) + \varphi'(\mathbf{S}_L)\left(\epsilon_L^{(n)}\mathbf{S}_L + \epsilon_R^{(n)}\mathbf{S}_R\right) + o\left(\left\|\epsilon^{(n)}\right\|\right) \tag{87}$$

where $\varphi'(\mathbf{Y})$ is a diagonal matrix of elements $[\varphi'(\mathbf{Y})]_{ii} = (\partial\varphi_i(Y_i)/\partial Y_i)$.

The correlation matrix $\mathbf{R}_{s,\varphi}$ at the extraction is shown in (88) at the bottom of page. The term $\mathrm{E}[\mathbf{S}_L\mathbf{S}_R^T\epsilon_R^{T(n)}\varphi'(\mathbf{S}_L)]$ vanishes in a first order approximation. This is due to the fact that $\epsilon_{ii}^{(n)}\ \forall i = 1, \ldots, P$ is an $o(\|\epsilon^{(n)}\|)$ term, while for $i \neq j$, the independence and zero mean assumptions for the sources enforce that

$$\left[\mathbf{E}\left[\mathbf{S}_L\mathbf{S}_R^T\epsilon_R^{T(n)}\varphi'(\mathbf{S}_L)\right]\right]_{ij} = \sum_{k=P+1}^{N}\mathbf{E}\left[S_i\varphi'(S_j)\right]\mathbf{E}[S_k]\epsilon_{jk}^{(n)} = 0 \tag{89}$$

Multiplying (88) by $\mathbf{D}\mathbf{G}^{(n)} = \mathbf{D}[\mathbf{I}_E + \epsilon_L^{(n)}, \epsilon_R^{(n)}]$ yields $\mathbf{R}_{s,\varphi}\mathbf{D}\mathbf{G}^{(n)}$. Substituting the result in (85) and, taking into account that $\mathrm{E}[\mathbf{S}_L\varphi^T(\mathbf{S}_L)]$ is a diagonal matrix, we obtain the left and right updates of the global system, respectively, as

$$\begin{aligned}\mathbf{G}_L^{(n+1)} &= \mathbf{G}_L^{(n)} - \mu\left(\mathbf{D}\mathbf{E}\left[\varphi'(\mathbf{S}_L)\epsilon_L^{(n)}\mathbf{S}_L\mathbf{S}_L^T\right]\right. \\ &\qquad \left. -\mathbf{E}\left[\mathbf{S}_L\varphi^T(\mathbf{S}_L)\right]\mathbf{D}\epsilon_L^{(n)}\right) \\ &\quad + \mu\left(\mathbf{D}\mathbf{E}\left[\varphi'(\mathbf{S}_L)\epsilon_L^{(n)}\mathbf{S}_L\mathbf{S}_L^T\right]\right. \\ &\qquad \left. -\mathbf{E}\left[\mathbf{S}_L\varphi^T(\mathbf{S}_L)\right]\mathbf{D}\epsilon_L^{(n)}\right)^T \\ &\quad + o\left(\left\|\epsilon^{(n)}\right\|\right) \end{aligned} \tag{90}$$

$$\begin{aligned}\mathbf{G}_R^{(n+1)} &= \mathbf{G}_R^{(n)} + \mu\left(\mathbf{E}\left[\mathbf{S}_L\varphi^T(\mathbf{S}_L)\right]\mathbf{D}\epsilon_R^{(n)}\right. \\ &\qquad \left. -\mathbf{D}\mathbf{E}\left[\varphi'(\mathbf{S}_L)\epsilon_R^{(n)}\mathbf{S}_R\mathbf{S}_R^T\right]\right) \\ &\quad + o\left(\left\|\epsilon^{(n)}\right\|\right). \end{aligned} \tag{91}$$

After truncating higher order terms $o(\|\epsilon^{(n)}\|)$ we rewrite the previous iterations in terms of the perturbation ϵ and of the stability factors $\kappa_i = \mathrm{E}[\varphi_i']\mathrm{Cov}(S_i) - \mathrm{E}[S_i\varphi_i]$. This gives the linearized dynamic of the algorithm around the extraction point

$$\epsilon_{ij}^{(n+1)} = \left(1 - \mu(d_i\kappa_i + d_j\kappa_j)\right)\epsilon_{ij}^{(n)} \tag{92}$$

$$\epsilon_{iq}^{(n+1)} = (1 - \mu d_i\kappa_i)\epsilon_{iq}^{(n)} \tag{93}$$

for $i, j|_{i\neq j} = 1, \ldots, P$; $q = P+1, \ldots, N$. 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 (61)–(63) for the algorithm step size.

E. Proof of Corollary 1

When the nonlinearities $\varphi_i(S_i)$ are the score functions of the densities of the sources, the definition of the stability factors κ_i take the form

$$\kappa_i = -\mathbf{E}\left[\frac{\partial^2 \log p_{S_i}(s_i)}{\partial s_i^2}\right]\mathrm{Cov}(S_i) + \mathbf{E}\left[S_i\frac{\partial \log p_{S_i}(s_i)}{\partial s_i}\right]. \tag{94}$$

The first expectation with the minus sign is the Fisher information of the distribution of S_i and can be rewritten as $\mathrm{E}[(\partial \log p_{S_i}(s_i)/\partial s_i)^2]$. Assuming the regularity condition $\lim_{s_i\to\infty} s_i p_{S_i}(s_i) = 0$ and integrating by parts, the second expectation in (94) simplifies to -1. Substituting both results yields the first desired expression

$$\kappa_i = \mathrm{Cov}(S_i)\mathbf{E}\left[\left(\frac{\partial \log p_{S_i}(s_i)}{\partial s_i}\right)^2\right] - 1. \tag{95}$$

$$\mathbf{E}\left[\mathbf{S}\varphi^T(\mathbf{Y})\right] = \begin{bmatrix}\mathbf{E}\left[\mathbf{S}_L\varphi^T(\mathbf{S}_L)\right] + \mathbf{E}\left[\mathbf{S}_L\mathbf{S}_L^T\epsilon_L^{T(n)}\varphi'(\mathbf{S}_L)\right] + \mathbf{E}\left[\mathbf{S}_L\mathbf{S}_R^T\epsilon_R^{T(n)}\varphi'(\mathbf{S}_L)\right] \\ \mathbf{E}\left[\mathbf{S}_R\mathbf{S}_R^T\epsilon_R^{T(n)}\varphi'(\mathbf{S}_L)\right]\end{bmatrix} + o\left(\left\|\epsilon^{(n)}\right\|\right). \tag{88}$$

The second part of the corollary results from the fact that for a Gaussian random variable $S_i^{\mathcal{N}}$, with the same mean and covariance of S_i, it is verified that

$$\begin{aligned}\mathrm{Cov}(S_i)\mathbf{E}&\left[\left(\frac{\partial \log p_{S_i^{\mathcal{N}}}(s_i)}{\partial s_i}\right)^2\right]\\ &=\mathrm{Cov}(S_i)\int\left(\frac{\partial \log p_{S_i^{\mathcal{N}}}(s_i)}{\partial s_i}\right)^2 p_{S_i}(s_i)ds_i\\ &=1 \end{aligned}\tag{96}$$

and that the following cross-term is

$$-2\mathrm{Cov}(S_i)\mathbf{E}\left[\left(\frac{\partial \log p_{S_i^{\mathcal{N}}}(s_i)}{\partial s_i}\right)\left(\frac{\partial \log p_{S_i}(s_i)}{\partial s_i}\right)\right]=-2. \tag{97}$$

Then, we can replace the constant -1 in (95) by the sum of (96) and (97). This completes the square

$$\kappa_i=\mathrm{Cov}(S_i)\mathbf{E}\left[\left(\frac{\partial}{\partial s_i}\log p_{S_i}(s_i)-\frac{\partial}{\partial s_i}\log p_{S_i^{\mathcal{N}}}(s_i)\right)^2\right] \tag{98}$$

and grouping the logarithms proves the second part of the corollary.

REFERENCES

[1] A. Cichocki and S.-I. Amari, *Adaptive Blind Signal and Image Processing*. New York: Wiley, 2002.

[2] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[3] R. A. Wiggins, "Minimum entropy deconvolution," *Geoexploration*, vol. 16, pp. 21–35, 1978.

[4] B. Godfrey, "An Information Theory Approach to Deconvolution," Stanford Exploration Project (SEP), Report no. 15, 1978.

[5] J. F. Claerbout, "Minimum Information Deconvolution," Stanford Exploration Project (SEP), Report no. 15, 1978.

[6] W. C. Gray, "Variable Norm Deconvolution," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1979.

[7] D. Donoho, *On Minimum Entropy Deconvolution*, D. F. Findley, Ed. New York: Academic, 1981, Appl. Time Ser. Anal. II, pp. 565–608.

[8] J. H. Friedman and J. W. Tukey, "A projection pursuit algorithm for exploratory data analysis," *IEEE Trans. Comput.*, vol. C-23, pp. 881–890, Sept. 1974.

[9] J. H. Friedman, W. Stuetzle, and A. Schroeder, "Projection pursuit density estimation," *J. Amer. Statist. Assoc.*, vol. 79, no. 387, pp. 599–608, Sept. 1984.

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

[11] J. H. Friedman, "Exploratory projection pursuit," *Amer. Statist. Assoc.*, vol. 82, no. 397, pp. 249–266, Mar. 1987.

[12] C. Jutten and J. Herault, "Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture," *Signal Process.*, vol. 24, pp. 1–10, 1991.

[13] J. F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 86, pp. 2009–2025, Oct. 1998.

[14] S. I. Amari and A. Cichocki, "Adaptive blind signal processing—neural network approaches," *Proc. IEEE*, vol. 86, pp. 2026–2048, Oct. 1998.

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

[16] X.-R. Cao and R. Liu, "General approach to blind source separation," *IEEE Trans. Signal Processing*, vol. 44, pp. 562–571, Mar. 1996.

[17] R. Linsker, "Self-organization in a perceptual network," *Comput.*, vol. 21, pp. 105–107, 1988.

[18] J. P. Nadal and N. Parga, "Non linear neurons in the low noise limit: A factorial code maximizes information transfer," *Network*, vol. 5, pp. 565–581, 1994.

[19] A. J. Bell and T. J. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computat.*, vol. 7, pp. 1129–1159, 1996.

[20] M. Girolami and C. Fyfe, *Negentropy and Kurtosis as Projection Pursuit Indices Provide Generalized ICA Algorithms*. Cambridge, MA: MIT Press, 1996, pp. 752–763.

[21] J. F. Cardoso, "Infomax and maximum likelihood for blind separation," *IEEE Signal Processing Lett.* , vol. 4, pp. 112–114, Apr. 1997.

[22] H. H. Yang and S. Amari, "Adaptive on-line learning algorithms for blind source separation—maximum entropy and minimum mutual information," *Neural Computat.*, vol. 9, no. 7, pp. 1457–1482, Oct. 1997.

[23] D. T. Pham and P. Garat, "Blind separation of mixture of independent sources through a quasimaximun likelihood approach," *IEEE Trans. Signal Processing*, vol. 45, pp. 1712–1725, July 1997.

[24] S. Amari and J. F. Cardoso, "Blind source separation—semiparametric statistical approach," *IEEE Trans. Signal Processing*, vol. 45, pp. 2692–2697, Nov. 1997.

[25] S. Amari, "Natural gradient work efficiently in learning," *Neural Computat.*, vol. 10, no. 2, pp. 251–276, 1998.

[26] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Netw.*, vol. 13, no. 4, pp. 411–430, 2000.

[27] ——, "A fast fixed-point algorithm for independent component analysis," *Neural Computat.*, vol. 9, pp. 1483–1492, 1997.

[28] S. Cruces, A. Cichocki, and S-i. Amari, *The Minimum Entropy and Cumulant Based Contrast Functions for Blind Source Extraction*. ser. Lecture Notes Comput. Sci., J. Mira and A. Prieto, Eds. New York: Springer-Verlag, 2001, pp. 786–793.

[29] J. F. Cardoso, "Entropic contrast for source separation: geometry and stability," in *Unsupervised Adaptive Filtering*, S. Haykin, Ed. New York: Wiley, 2000, vol. I, pp. 139–190.

[30] J. Principe, D. Xu, and J. Fisher, "Information theoretic learning," in *Unsupervised Adaptive Filtering*, S.Simon Haykin, Ed. New York: Wiley, 2000, vol. I, pp. 265–319.

[31] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[32] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 105, no. 4, pp. 620–630, 1957.

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

[34] S. Cruces, A. Cichocki, and S. Amari, "On a new blind signal extraction algorithm: different criteria and stability analysis," *IEEE Signal Processing Lett.*, vol. 9, pp. 233–236, Aug. 2002.

[35] A. Cichocki, R. Thawonmas, and S. Amari, "Sequential blind signal extraction in order specified by stochastic properties," *Electron. Lett.*, vol. 33, no. 1, pp. 64–65, 1997.

[36] 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, ch. 3.

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

[38] A. Edelman, T. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM J. Matrix Anal. Appl.*, vol. 20, pp. 303–353, 1998.

[39] S. Amari, "Natural gradient learning for over- and under-complete bases in ica," *Neural Computat.*, vol. 11, pp. 1875–1883, 1999.

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

[41] A. Cichocki, S. Amari, and K. Siwek *et al.*. (2003) ICALAB Toolboxes. [Online]URL: http://www.bsp.brain.riken.go.jp/ICALAB

Sergio A. Cruces-Alvarez (S'94–A'99) was born in Vigo, Spain, in 1970. He received the Telecommunication Engineer and Ph.D. degrees from the University of Vigo, Spain, in 1994 and 1999, respectively.

From 1994 to 1995, he worked as a Project Engineer for the Department of Signal Theory and Communications, University of Vigo. He has been a Visitor at the Laboratory for Advanced Brain Signal Processing under the Frontier Research Program Riken, Japan, on several occasions. He is currently an Associate Professor at the University of Seville, Spain, where he has been a member of Signal Theory and Communications Group since 1995. He teaches undergraduate and graduate courses on digital signal processing of speech signals and mathematical methods for communication. His current research interests include statistical signal processing, information theoretic and neural network approaches, blind equalization, and filter stabilization techniques.

Andrzej Cichocki (M'96) was born in Poland. He received the M.Sc. (with honors), Ph.D., and Habilitate Doctorate (Doctor of Science) degrees, all in electrical engineering, from Warsaw University of Technology, Warsaw, Poland, in 1972, 1975, and 1982, respectively.

Since 1972, he has been with the Institute of Theory of Electrical Engineering, Measurements and Information Systems, Warsaw University of Technology, where he became a Full Professor in 1991. He spent a few years at the 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 Head of Laboratory for Advanced Brain Signal Processing. He is the coauthor of three books: MOS Switched-Capacitor and Continuous-Time Integrated Circuits and Systems (New York: Springer-Verlag, 1989), Neural Networks for Optimization and Signal Processing (New York: Teubner-Wiley, 1993/94), and Adaptive Blind Signal and Image Processing (New York: Wiley, 2003) and more than 150 research journal papers. His current research interests include signal and image processing, especially analysis and processing of multi-sensory, and multimodal biomedical data.

Prof. Cichocki is a Member of the IEEE SP Technical Committee for Machine Learning for Signal Processing and the IEEE Circuits and Systems Technical Committee for Blind Signal Processing.

Shun-ichi Amari (S'71–M'88–SM'92–F'94) was born in Tokyo, Japan in 1936. He received the B.Sc. and Dr. Eng. degrees in mathematical engineering from the Department of Applied Physics, University of Tokyo, in 1958 and 1963, respectively.

He was an Associate Professor in the Department of Communication Engineering, Kyushu University from 1963 to 1967. Later, he returned to the Department of Mathematical Engineering and Information Physics, University of Tokyo, where he retired as a Full Professor in 1996 and then a Professor Emeritus. Since 1994, he has been with the Brain Science Institute, Riken, Japan where he leads his own laboratory in mathematical neuroscience and he also manages the institute as its Director. He is the Founding Co-Editor of *Neural Networks* and *Applied Mathemtaics*. He has made a considerable number of significant contributions to the mathematical foundations of neural network theory. In the late 1970s, he initiated a new approach of information science, called "Information Geometry." His approach brought together new concepts in modern differential geometry, which is connected with statistics, information theory, control theory, general learning theory, pattern recognition, and independent component analysis. The applications of information geometry spread over many areas of information sciences, including brain science and artificial intelligence.

Prof. Amari was a Member of the Japanese Council of Scientists from 1997 to 2000, President of the International Neural Network Society in 1989, Council Member of the Bernoulli Society for Mathematical Statistics and Probability Theory from 1995 to 1999, and Vice President of the IEICE from 1995 to 1997.