A Minimum-Range Approach to Blind Extraction of Bounded Sources

Frédéric Vrins, *Student Member, IEEE*, John A. Lee, and Michel Verleysen, *Senior Member, IEEE*

***Abstract*—In spite of the numerous approaches that have been derived for solving the independent component analysis (ICA) problem, it is still interesting to develop new methods when, among other reasons, specific *a priori* knowledge may help to further improve the separation performances. In this paper, the minimum-range approach to blind extraction of bounded source is investigated. The relationship with other existing well-known criteria is established. It is proved that the minimum-range approach is a contrast, and that the criterion is discriminant in the sense that it is free of spurious maxima. The practical issues are also discussed, and a range measure estimation is proposed based on the order statistics. An algorithm for contrast maximization over the group of special orthogonal matrices is proposed. Simulation results illustrate the performances of the algorithm when using the proposed range estimation criterion.**

***Index Terms*—Blind source separation (BSS), bounded sources, discriminacy, independent component analysis (ICA), order statistics, range estimation, Stiefel manifold.**

I. Introduction

INDEPENDENT component analysis (ICA) [1]–[3] has received some attention for more than two decades, due to its numerous applications in multichannel signal processing, especially in biomedical signal processing, seismic signal analysis, denoising in electric and magnetic circuits, and image processing.

Many ICA algorithms based on various objective functions have been derived to achieve blind source separation (BSS), either by extracting the sources one by one (deflation approach), or by separating all the sources at once (simultaneous approach). Among others, we can cite JADE [4], FastICA [5], EFICA [6], Infomax [7], extended Infomax [8], RADICAL [9], MISEP [10], or nonparametric ICA [11]. For a detailed review, we refer the reader to the monograph by Cichocki and Amari [12]. ICA algorithms may perform differently, depending on the kind of sources that are involved in the mixtures. Most ICA researchers agree that there does not exist a "super-algorithm," making all other ICA approaches useless; new approaches still arise by using prior information, such as, e.g., sparsity or nonnegativity [13], [14] or still other constrains [15].

Manuscript received January 27, 2006; revised July 10, 2006 and October 25, 2006; accepted November 2, 2006.

F. Vrins and M. Verleysen are with the Microelectronics Lab (DICE), Université catholique de Louvain, Louvain-la-Neuve 1348, Belgium (e-mail: vrins@dice.ucl.ac.be; verleysen@dice.ucl.ac.be).

J. A. Lee is with the Molecular Imaging and Experimental Radiotherapy Department, Saint-Luc University Hospital, Université catholique de Louvain, Brussels 1200, Belgium (e-mail: John.Lee@imre.ucl.ac.be).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNN.2006.889941

An important issue in adaptive techniques is the problem of "false maxima" of the objective function. In the BSS framework, this problem has been proved to exist for some specific criteria such as, e.g., Shannon's entropy or mutual information [11], [16], [17]: locally maximizing these criteria is not equivalent to recover the sources. Therefore, several criteria that do not suffer from spurious maxima have been developed. For example, under the whitening constraint, the local maximum points of the square of the output kurtosis used in a deflation scheme correspond to the extraction of the sources [18] (historically, the spurious maxima problem was the motivation that has yielded the deflation method). This is also proved for the simultaneous case when two sources are involved in the mixtures [19] (the proof is not extended to a higher number of sources, but experimental results illustrate the good behavior of the criterion); the two-sources BSS problem reduces to phase estimation. In the same order of idea, the limit points of geometric ICA are shown to be the solutions of the BSS problem, at least for two-sources and for symmetric, unimodal densities [20].

In this paper, the sources are assumed to be bounded, i.e., the source support measure is finite. This assumption has yielded different approaches to solve the BSS problem using simultaneous techniques (based on geometrical or statistical methods [21]–[23]). More recently, deflation approaches have been independently proposed in [24] and [25] based on information-theory and statistical properties, respectively; both use support-driven information: the support measure itself or the measure of its convex hull (also known as the "range"). If the support is not convex, support measure and range may be different; as an example, if X is a random variable with support $\Omega(X) = [-2,1] \cup [2,5] \setminus \{3\}$, then its measure equals six but its convex hull is $\bar{\Omega}(X) = [-2,5]$ and the range of X, which is the measure of this convex hull, is $R(X) = 7$. The wide variety of techniques tailored for bounded sources stems from the following facts: 1) bounded sources are often encountered in practice (e.g., digital images, communication signals, and analog electric signals varying within the range of power voltage) and 2) simple and powerful BSS methods can be derived in this specific context.

We focus here on a deflation method; the BSS problem is referred to as blind extraction of bounded sources (BEBS). This work presents an extension of [25]; additionally to [24], it covers both the complete theoretical analysis of the extreme points (including the spurious optima) of the range-based criterion and the practical issues related to its estimation for an arbitrary number of sources. In [25], most of the proofs were only sketched, and the practical aspects were not discussed. The paper is organized as follows. First, a specific contrast for mixtures of bounded sources is derived in Section II. Next, relationships with mutual

information, negentropy, Renyi's entropy, and kurtosis-based approaches are emphasized in Section III, before proving several properties of the criterion in Section IV. One of the main result is the so-called *discriminacy*, which states that each local maximum of the contrast function corresponds to a satisfactory solution of the BEBS problem. In Section V, a finite-sample estimator of the support convex hull measure is proposed for the contrast, based on averaged order statistic differences, i.e., averaged quasi-ranges (other ICA methods also use order statistics for density, quantile, or distribution functions estimation (see, e.g., [23], [26], [27], and references therein). A batch algorithm is provided in Section VI for the contrast maximization. Simulation results illustrate the good performances of the method. The proofs are relegated to the Appendices I–V.

II. Mixture Model and Proposed Contrast

Within the ICA framework, BSS aims at separating independent zero-mean source signals $\mathbf{S}(t) = [S_1(t), \cdots, S_K(t)]^T$ from K linear instantaneous mixtures of them $\mathbf{X}(t) = [X_1(t), \cdots, X_K(t)]^T$

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

where $\mathbf{A}$ is the so-called mixing matrix of order K, thus assumed to be square and nonsingular. The source signals are supposed to be white [1]: the source covariance matrix is obviously diagonal because of the independence assumption, and the magnitude of the ith source S_i can be divided by $\sqrt{\mathrm{Var}(S_i)}$ without changing the mixture model given in (1), provided that the ith column of $\mathbf{A}$ is multiplied by the same scaling coefficient. Then, the target of BSS is now to recover independent unit-variance signals. It is also assumed in the following that the source densities are constant in time (so that t shall be omitted in the equations).

ICA aims at finding a separating (also called unmixing) matrix $\mathbf{B}$ such that the independence between the outputs $\mathbf{Y} = \mathbf{B}\mathbf{X}$ is maximized with $\mathbf{Y} = [Y_1, \cdots, Y_K]^T$.

A. Whiteness and Stiefel Manifold

Most ICA approaches assume prewhitening, that is $E[\mathbf{X}\mathbf{X}^T] = \mathbf{I}_K$ (where $\mathbf{I}_K$ is the identity matrix of order K). If it is not the case, we can simply premultiply the mixtures by a whitening matrix $\mathbf{V}$: $\mathbf{X} \leftarrow \mathbf{V}\mathbf{X}$ implying that the mixing matrix is now $\mathbf{A} \leftarrow \mathbf{V}\mathbf{A}$. Matrix $\mathbf{V}$ can easily be obtained by eigenvalue decomposition of the mixtures covariance matrix [1].

If $\mathbf{X}$ is white, $\mathbf{A}$ can be assumed to be in the group $\mathcal{O}(K)$ of the $K \times K$ orthogonal matrices. Clearly, since we are searching for uncorrelated sources, any satisfactory unmixing matrix should satisfy $\mathbf{b}_i\mathbf{b}_j^T \propto \delta_{i,j}$, where $\mathbf{b}_i$ denotes the ith row of $\mathbf{B}$ and $\delta_{i,j}$ is the Kronecker delta.

Because whiteness is preserved only under orthogonal transformations, one can restrict the search to the set of unmixing matrices $\mathbf{B} \in \mathcal{O}(K)$. The orthogonal group of order K forms a $K(K-1)/2$-dimensional subspace of $\mathbb{R}^{K\times K}$, called *Stiefel* manifold [2], [28].[1] More specifically, since one can only identify $\mathbf{A}^{-1}$ up to a left multiplication by the product of gain and permutation matrices [2], we can also freely assume that $\mathbf{B} \in \mathcal{SO}(K)$, the group of orthogonal matrices with $\det \mathbf{B} = 1$ without adding further indeterminacies.

The set of target unmixing matrices (corresponding to satisfactory solution of the BSS problem) can be defined as $\mathcal{B}^\star \triangleq \{\mathbf{B} : \mathbf{B} = \mathbf{\Lambda}\mathbf{P}\mathbf{A}^{-1}\}$, where $\mathbf{\Lambda}$ and $\mathbf{P}$ can be any diagonal and permutation nonsingular matrices in $\mathbb{R}^{K\times K}$, respectively. Note that obviously $\mathcal{SO}(K) \cap \mathcal{B}^\star \neq \emptyset$. For convenience, we define the global transfer matrix $\mathbf{C}$ as $\mathbf{C} \triangleq \mathbf{B}\mathbf{A}$.

B. Deflation Criterion for BEBS

A particular contrast for BEBS can be built. Let us denote $\Omega(\mathbf{b}_i\mathbf{X})$ the support of $Y_i = \mathbf{b}_i\mathbf{X}$, that is the set where the probability density function (pdf) of Y_i is strictly positive, and $\bar{\Omega}(.)$ denotes the smallest convex set including $\Omega(.)$. We define the range of a random variable X as

$$R(X) \triangleq \mu\left[\bar{\Omega}(X)\right] \tag{2}$$

where $\mu[.]$ is the (Lebesgue) measure of sets, which is the interval length in the one-dimensional (1-D) case. Then, we shall prove in Section IV that the following criterion is a contrast for BEBS:

$$\mathcal{C}(\mathbf{b}_i) \triangleq -R(\mathbf{b}_i\mathbf{X})/\sqrt{\mathrm{Var}(\mathbf{b}_i\mathbf{X})}. \tag{3}$$

If the criterion is maximized subject to $\mathrm{Var}(\mathbf{b}_i\mathbf{X}) = \mathbf{b}_i\mathbf{b}_i^T = \|\mathbf{b}_i\|^2 =$constant, the denominator can be omitted.

Though the contrast property of $\mathcal{C}(\mathbf{b}_i)$ will be rigorously proved in Section IV, we now show how the last criterion can be obtained.

Let us first observe that $\bar{\Omega}(.)$ is a simple interval, and thus $R(\alpha U) = |\alpha| R(U)$, for all $\alpha \in \mathbb{R}$. Furthermore, one has

$$\begin{aligned} R(U+V) &= \sup\bar{\Omega}(U+V) - \inf\bar{\Omega}(U+V) \\ &= \sup\bar{\Omega}(U) + \sup\bar{\Omega}(V) \\ &\quad - \left(\inf\bar{\Omega}(U) + \inf\bar{\Omega}(V)\right) \\ &= R(U) + R(V) \end{aligned} \tag{4}$$

where $\Omega(U+V) = \{u+v | u \in \Omega(U), v \in \Omega(V)\}$, in which U and V are independent random variables. Hence, noting by $[\mathbf{C}]_{ij}$ the element i,j of matrix $\mathbf{B}\mathbf{A}$

$$\begin{aligned} R(\mathbf{b}_i\mathbf{X}) = \sum_{j=1}^{K} R\left([\mathbf{C}]_{ij}S_j\right) &= \sum_{j=1}^{K} |[\mathbf{C}]_{ij}|\, R(S_j) \\ &= |\mathbf{b}_i\mathbf{A}|.R(\mathbf{S}) \end{aligned} \tag{5}$$

where $R(\mathbf{S}) = [R(S_1), \cdots, R(S_K)]^T$ and the absolute value is element-wise.

It can be intuitively understood by looking at (5) that $\mathcal{C}(\mathbf{b}_i)$ is a contrast for deflation-based ICA. Indeed, this criterion can be written as (3) and is thus not sensitive to the scale of Y_i; if we constrain $\|\mathbf{b}_i\| = 1$, the criterion is maximized when $\mathbf{b}_i\mathbf{X}$ is proportional to a source with the smallest range. In practice, the range must be estimated from a finite number of samples, so that one is led to maximize a finite-sample approximation $\hat{\mathcal{C}}(\mathbf{b}_i)$ of $\mathcal{C}(\mathbf{b}_i)$. For example, a simple approximation

[1] In this paper, we reduce the Stiefel manifold to the group of orthogonal matrices, even if it holds more generally for rectangular matrices, too.

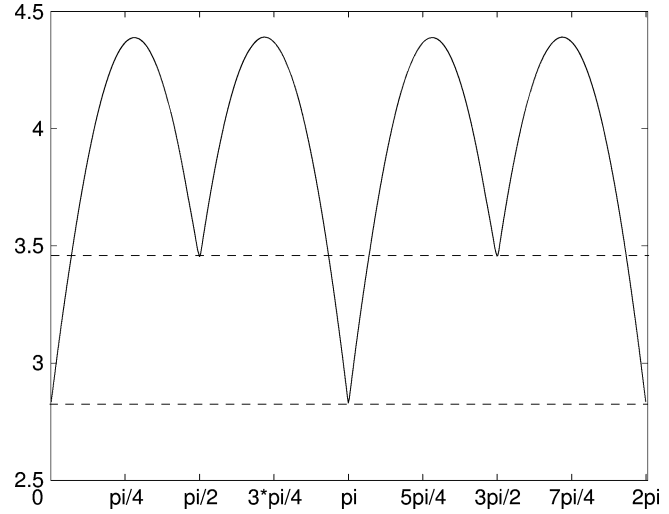

Fig. 1. Minimum range criterion: evolution of $R(Y_1)$ with respect to θ. The 1000 samples unit-variance source signals are a sine wave (S_1) and a random signal with uniform distribution (S_2); the convex source supports have measures equal to $R(S_1) = 2\sqrt{2}$ and $R(S_2) = 2\sqrt{3}$, respectively.

of $R(\mathbf{b}_i\mathbf{X})$ would be the empirical range of $\mathbf{b}_i\mathbf{X}$, defined as $R^\star(\mathbf{b}_i\mathbf{X}) \triangleq \max_{t_1,t_2}[\mathbf{b}_i\mathbf{X}(t_1) - \mathbf{b}_i\mathbf{X}(t_2)]$. Further, in order that, generally speaking, the estimated range of a sum equals the sum of the estimated ranges, it is needed that some specific sample points are observed. For instance, each of the K sources have to reach simultaneously their maximum value at a same time $t_{\max}$, and likewise for the minimum, it must exist a time index $t_{\min}$ such that each of the source reaches its minimum value at $t = t_{\min}$. In this case, $R^\star(\sum_{j=1}^K [\mathbf{C}]_{ij}\mathbf{S}_j) = R^\star(\mathbf{b}_i\mathbf{X}) = \mathbf{b}_i\mathbf{X}(t_{\max}) - \mathbf{b}_i\mathbf{X}(t_{\min}) = \sum_{j=1}^K [\mathbf{C}]_{ij}S_j(t_{\max}) - \sum_{j=1}^K [\mathbf{C}]_{ij}S_j(t_{\min}) = \sum_{j=1}^K R^\star([\mathbf{C}]_{ij}S_j)$. This is clearly the case if the sources are independent when the sample set is large enough.

C. Contrast Interpretation

The geometrical interpretation of the minimum output range used in a deflation approach to ICA is straightforward. Assuming that $p-1$ sources have already been recovered, a pth source can be extracted by searching direction $\mathbf{b}_p$ orthogonal to the subspace spanned by $\mathbf{b}_1, \cdots, \mathbf{b}_{p-1}$ such that the projection of the output pdf onto $\mathbf{b}_p$ has a minimum range with unit variance. Fig. 1 shows the output range of $Y_1 = [\cos\theta, \sin\theta][S_1, S_2]^T$ as a function of the transfer angle θ. Section III will be dedicated to establishing the relationships between the range and Renyi's entropy, as well as with the kurtosis. In the remaining part of the paper, several properties will be shown, proving rigorously the previous intuitive result, as well as the so-called discriminacy property of $\mathcal{C}(\mathbf{b}_i)$. Next, an optimization algorithm will be provided for this contrast, which is not continuously differentiable everywhere.

III. Relationship to Other ICA Approaches

Most often, in order to make $\mathbf{B}$ converge from an initial point in $\mathbb{R}^{K\times K}$ to $\mathbf{B} \in \mathcal{B}^\star$, a so-called *contrast function* f is maximized by an adaptive method, f reflecting the statistical independence between the Y_i. One of the most known contrast function is

$$\tilde{\mathcal{C}}(\mathbf{B}) = \log|\det\mathbf{B}| - \sum_{i=1}^K H(\mathbf{b}_i\mathbf{X}) \tag{6}$$

where $H(X) = -\int f_X(x)\log f_X(x)dx$ is Shannon's entropy. Maximizing $\tilde{\mathcal{C}}(\mathbf{B})$ yields $\mathbf{B} \in \mathcal{B}^\star$. Observe that the $\log|\det\mathbf{B}|$ term vanishes if $\mathbf{B}$ is constrained at each step to belong to any subset of $\mathcal{O}(K)$.

A. Symmetric Approach to Minimum Range ICA

Several years ago, Pham proposed to replace the functional $H(.)$ by the range $R(.)$ [23]; he proved that if the sources are bounded, then $-\log\{\prod_{i=1}^K R(\mathbf{b}_i\mathbf{X})/|\det\mathbf{B}|\}$ is a contrast for the simultaneous separation of the sources.

The relationship between this simultaneous criterion and the $\mathcal{C}(\mathbf{b}_i)$, $1 \le i \le K$ is now obvious: Pham's criterion corresponds to summing the log of the $-1/\mathcal{C}(\mathbf{b}_i)$, when the criterion is optimized over $\mathcal{O}(K)$ or one of its subsets.

It is explained in [23] that maximizing $-\log\{\prod_{i=1}^K R(\mathbf{b}_i\mathbf{X})/|\det\mathbf{B}|\}$ amounts at looking for a "hyper-parallelepiped" with smallest volume enclosing the support $\Omega(\mathbf{BX})$.

B. Relationship to Renyi's Entropy

The r-order Renyi's entropy, $H_r(X)$ is defined as [29], [30]

$$\begin{cases} \frac{1}{1-r}\log\left\{\int f_X^r(x)dx\right\}, & \text{for } r\in\{[0,1)\cup(1,\infty)\} \\ H(X) = -\int f_X(x)\log f_X(x)dx, & \text{for } r=1. \end{cases} \tag{7}$$

Note that the integration set in the previous integrals is the support $\Omega(X)$ of X. Obviously, $H_0(X) = \log\mu[\Omega(X)]$ [31]. This was also pointed out in the BSS framework by Cruces and Duran [24].

If we consider a modified zero-Renyi's entropy $\bar{H}_0(X)$ such that the integration domain in (7) is extended to the convex hull of the support of X, then $\mu[\bar{\Omega}(X)] = \exp(\bar{H}_0(X))$ [24], [32].

Then, minimizing $\bar{H}_0(\mathbf{b}_i\mathbf{X})$ with respect to $\mathbf{b}_i$ is equivalent to finding the vector $\mathbf{b}_i$ such that the volume of the convex hull of $\Omega(\mathbf{b}_i\mathbf{X})$, i.e. $R(\mathbf{b}_i\mathbf{X})$, is minimum.

C. Relationship to Absolute Kurtosis

The expression of the output minimum range criterion in (5) is similar to the output absolute kurtosis $|\kappa(Y_i)|$, another contrast for ICA. Recall that the kurtosis of any zero-mean and unit-variance random variable can be written as [1]

$$\kappa(Y_i) = E\left[Y_i^4\right] - 3 \tag{8}$$

where

$$\kappa(Y_i) = \sum_{j=1}^K [\mathbf{C}]_{ij}^4 \kappa(S_j). \tag{9}$$

In both (5) and (9), the criteria can be decomposed as the dot product between a vector of positive functions of the mixing weights and a vector of mappings of the source densities. Both the range and the kurtosis have the form

$$\phi(\mathbf{c}_i).\psi^T(\mathbf{S}) \tag{10}$$

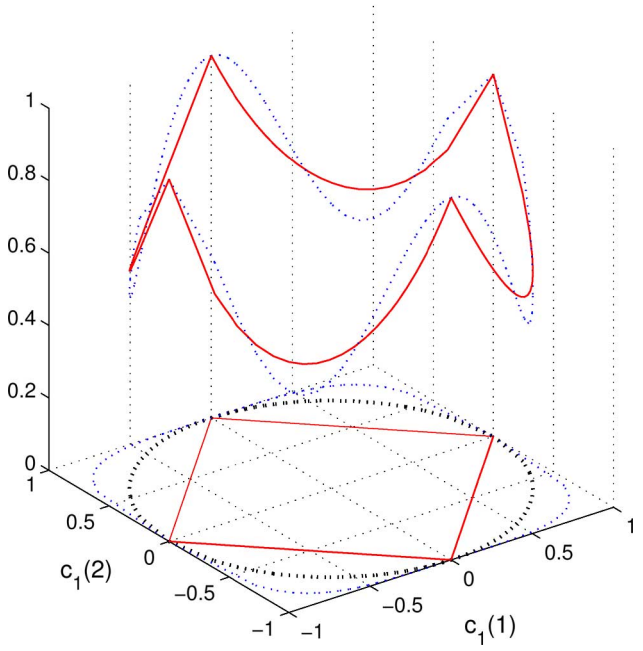

Fig. 2. Kurtosis and output range landscapes. The real functions $\kappa(Y_1)$ (dot) and $2 - R(Y_1)$ (solid) are plotted on the $\|\mathbf{c}_1\|$ 1-D manifold, that is on the unit circle. The maximum values are obtained for $\mathbf{c}_1 \in \{\pm\mathbf{e}_1, \pm\mathbf{e}_2\}$. In the horizontal plane $Z = 0$, the highest isolevel curves of the contrasts having nonempty intersection with the circle constraint (dash) are plotted. This intersection reduces to the set $\{\pm\mathbf{e}_1, \pm\mathbf{e}_2\}$.

where $\phi(\mathbf{w}) \triangleq [\phi(\mathbf{w}(1)), \cdots, \phi(\mathbf{w}(K))]$ and $\psi(\mathbf{S}) \triangleq [\psi(S_1), \cdots, \psi(S_K)]$.

For the range criterion, $\phi(.)$ is the absolute value function $|.|$ and $\psi(.)$ is the source range $R(S_i)$, as shown by (5). For the kurtosis criterion, $\phi(.)$ is the fourth power and $\psi(.)$ is the source kurtosis. Note that to deal simultaneously with both negative-kurtosis and positive-kurtosis sources, the absolute value of $\kappa(Y_i)$ is often considered in BSS application. As the range is always positive, both absolute kurtosis-based and range-based contrast functions share the form $|\phi(\mathbf{c}_i).\psi^T(\mathbf{S})|$.

For illustration purposes, suppose that $K = 2$, both source kurtoses and ranges equal to one, and $\|\mathbf{c}_i\| = 1$. Clearly, we can set $\mathbf{c}_i = [\cos\theta, \sin\theta]$. Noting that $Y_i = \mathbf{c}_i\mathbf{S}$, we have $R(Y_i) = |\cos\theta| + |\sin\theta|$ and $|\kappa(Y_i)| = \cos^4\theta + \sin^4\theta$. The first criterion is always greater than one (with equality if and only if $\theta \in \{k\pi/2, k \in \mathbb{Z}\}$) and the second is always lower than one (with equality if and only if $\theta \in \{k\pi/2, k \in \mathbb{Z}\}$).

The optimization landscapes of kurtosis and output range are similar. It is shown in Fig. 2 that the largest level curves of $\kappa(Y_1)$ and $2 - R(Y_1)$ intersect the $\|\mathbf{c}_1\| = 1$ constraint when $\mathbf{c}_1 \in \{\pm\mathbf{e}_1, \pm\mathbf{e}_2\}$, where the canonical vectors are defined as $\mathbf{e}_i(j) = \delta_{i,j}$. Basically, these two isolevel curves of the kurtosis and output range contrast functions correspond to the fourth power of the four-norm and the one-norm of $\mathbf{c}_i$ subject to the second-norm is kept unitary if all $\kappa(S_i) = 1$ or all $R(S_i) = 1$, respectively.

IV. Deflation Contrast Properties

Fig. 1 suggests three properties, at least in the $K = 2$ case. First, $R(Y_i)$ reaches its global minimum when $Y_i = \pm S_1$, where S_1 is the source with the lowest range. Second, a local minimum is obtained for $R(Y_i)$ when $Y_i = \pm S_j, j \in \{1, 2\}$. Third, no local minimum exists if $\theta \notin \{k\pi/2 | k \in \mathbb{Z}\}$. This section presents the formal derivation of the aforementioned properties of $\mathcal{C}(\mathbf{b}_i)$ for the general $K \geq 2$ case, which were first sketched in [25].

In the following, we will work directly on $f(\mathbf{c}_i) \triangleq \mathcal{C}(\mathbf{b}_i)$, where $\mathbf{c}_i = \mathbf{b}_i\mathbf{A}$ for simplifying as much as possible the following developments and notations (note that proving results in the transfer matrix space rather than in the unmixing matrix space does not matter here, as explained in Remark 1). Consequently, by looking at (5), it is obvious that whatever is $\mathbf{w} \in \mathbb{R}^K$, $f(\mathbf{w})$ is not sensitive to the sign of the K elements $\mathbf{w}(j)$ of the vector argument. It will be shown (see $\mathcal{A}_1$) that $\|\mathbf{w}\|$ has no impact on $f(\mathbf{w})$, too. Hence, the study of $\mathcal{C}(\mathbf{b}_i)$ can be restricted to the study of $f(\mathbf{w})$ with vectors $\mathbf{w} \in \mathcal{V}_K^\lambda \subset \mathbb{R}^K$, where

$$\mathcal{V}_K^\lambda \triangleq \left\{\mathbf{w} \in \mathbb{R}^K \text{s.t.} \|\mathbf{w}\| = \lambda, \mathbf{w}(j) > 0 \,\forall\, 1 \leq j \leq K\right\}.$$

Observe that $\mathcal{V}_K^\lambda$ is nothing else than the set of K-dimensional vectors of Euclidean norm equal to λ with positive entries. It can be interpreted as the intersection of R_+^K with the surface of the K-dimensional hypersphere centered at the origin with radius λ.

Remark 1: (Accessibility from f to $\mathcal{C}$) Fortunately, under the only constraint that $\|\mathbf{b}_i\|$ is fixed (implying $\|\mathbf{c}_i\| = \|\mathbf{b}_i\|$ by orthogonality of $\mathbf{A}$), one can freely adjust the $[\mathbf{C}]_{ij}$, even if such updates must be done by making $\mathbf{b}_i$ varying. In order to extract the pth source, one has to update $\mathbf{c}_p = \mathbf{b}_p\mathbf{A}$. However, since the columns of $\mathbf{A}$ form an orthonormal basis in $\mathbb{R}^{K\times K}$, any row vector $\mathbf{c}_p$ can be obtained by choosing an appropriate $\mathbf{b}_p$, which is orthogonal to $\mathbf{b}_1, \ldots, \mathbf{b}_{p-1}$. Hence, all propositions and theorems given below remain valid despite the fact that the transfer matrix elements must be updated through $\mathbf{B}$. Therefore, if f is a contrast, then so is $\mathcal{C}$.

A. Contrast Properties

The properties of a deflation contrast can be extended from the properties of a simultaneous contrast according to Comon [3]. In the remainder of this paper, we suppose that the following assumption holds, without loss of generality:

- $\mathcal{A}_0$: *Source ordering*. The sources are bounded and they are ordered by decreasing values of the contrast, that is $0 < R(S_1) = \ldots = R(S_k) < R(S_{k+1}) \leq \cdots \leq R(S_K) < \infty$. It is assumed that the k first sources have the same range, $1 \leq k \leq K$.

The mapping $f : \mathbb{R}^K \to \mathbb{R}$ is a deflation contrast if it satisfies the three following properties.

- $\mathcal{A}_1$: *Scaling invariance*. $f(\mathbf{w}) = f(\lambda\mathbf{w})$ for all $\lambda \in \mathbb{R}_0$.
- $\mathcal{A}_2$: *Global maximum*. The global maximum of $f(\mathbf{w}), \mathbf{w} \in \mathcal{V}_K^1$, is obtained when one of the first k sources is recovered, i.e., for $\mathbf{w} \in \{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$.
- $\mathcal{A}_3$: *Complete extraction*. Assuming that the $p - 1$ first sources have already been extracted, the global maximum of $f(\mathbf{w})$ subject to $\mathbf{w} \in \mathcal{V}_K^1$ and $\mathbf{w}\mathbf{c}_r^T = 0$ for all $1 \leq r < p$ is obtained for $\mathbf{w} \in \{\mathbf{e}_i : f(\mathbf{e}_i) = f(\mathbf{e}_p)\}$.

We will further show that f is a discriminant contrast, i.e., all the local maxima of the contrast are relevant for source separation as follows.

- $\mathcal{A}_4$: *Discriminacy property*. The set of local maximum points of $f(\mathbf{w})$, $\mathbf{w} \in \mathcal{V}_K^1$ reduces to $\{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$.

This is an interesting property, also shared by the sum of the output squared kurtosis contrast under the whitening constraint [18]. It gives confidence in solutions obtained by using gradient–ascent methods since there is no spurious maximum in the contrast function.

The four aforementioned properties $\mathcal{A}_1$–$\mathcal{A}_4$ will be proved in Sections IV-B–IV-E.

B. Scaling Invariance: $\mathcal{A}_1$

Obviously, when constraining $\mathbf{B} \in \mathcal{O}(K)$, the scaling problem is avoided. Without the $\|\mathbf{b}_i\| = 1$ constraint, the contrast $f(\mathbf{c}_i)$ becomes $-R(\mathbf{c}_i\mathbf{S})/\sqrt{\mathrm{Var}(Y_i)}$, which does not depend of the magnitude of Y_i. This proves $\mathcal{A}_1$.

C. Global Maximum: $\mathcal{A}_2$

Theorem 1, proved in Appendix I, shows that if $\mathbf{w} \in \mathcal{V}_K^\lambda$, then the global maximum of $f(\mathbf{w})$ corresponds to the extraction of one of the $k \leq K$ sources with the lowest range. This point is also mentioned in [24].

Theorem 1 (Global Maximum): Suppose that $\mathcal{A}_0$ holds. Then, one gets

$$\arg\max_{\mathbf{w} \in \mathcal{V}_K^\lambda} f(\mathbf{w}) \in \{\lambda.\mathbf{e}_1, \ldots, \lambda.\mathbf{e}_k\}.$$

Theorem 1 guarantees that f satisfies $\mathcal{A}_2$. The possible existence of local maxima is addressed in Sections IV-D and IV-E.

D. Complete Extraction: $\mathcal{A}_3$

Because of $\mathcal{A}_1$, one can restrict the analysis of $f(\mathbf{w})$ to $\mathbf{w} \in \mathcal{V}_K^1$ even though the mathematical developments can be easily extended to other values of λ.

Theorem 2 (Subset of Local Maximum Points): Function $f(\mathbf{w})$, subject to $\mathbf{w} \in \mathcal{V}_K^1$, admits a local maximum for $\mathbf{w} = \mathbf{e}_i$, $1 \leq i \leq K$.

Consider two vectors $\mathbf{p} \in \mathcal{V}_K^1$, $\mathbf{q} \in \mathcal{V}_K^1$, and let us introduce the associate contrast difference $\Delta f(\mathbf{p}, \mathbf{q})$ defined as

$$\Delta f(\mathbf{p}, \mathbf{q}) \triangleq f(\mathbf{p}) - f(\mathbf{q}). \tag{11}$$

The proof, detailed in Appendix II, shows that for any $\hat{\mathbf{e}}_i \in \mathcal{V}_K^1$ sufficiently close (but different) from $\mathbf{e}_i$, then $\Delta f(\mathbf{e}_i, \hat{\mathbf{e}}_i) > 0$.

Corollary 1 (Complete Extraction): Function $f(\mathbf{w})$, $\mathbf{w} \in \mathcal{V}_K^1$, satisfies $\mathcal{A}_3$.

By Theorem 2, we know that $f(\mathbf{w})$ subject to $\mathbf{w} \in \mathcal{V}_K^1$ reaches a local maximum if $\mathbf{w} \in \{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$. Then, assuming that the first $p-1$ sources have already been extracted, a pth source can be found by updating $\mathbf{c}_p$ where $\mathbf{c}_p(1) = \ldots = \mathbf{c}_p(p-1) = 0$. Next, discarding the $p-1$ first sources and setting $K \leftarrow K - p + 1$, Theorem 1 is used to prove that the global maximum of $f(\mathbf{w})$ and $\mathbf{w} \in \mathcal{V}_K^1$ equals now $f(\mathbf{e}_p)$ and is reached for $\mathbf{w} \in \{\mathbf{e}_i : f(\mathbf{e}_i) = f(\mathbf{e}_p)\}$, $p \leq i \leq K$.

E. Discriminacy Property: $\mathcal{A}_4$

The previous sections prove that $f(\mathbf{c}_i) = \mathcal{C}(\mathbf{b}_i)$ satisfy $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{A}_3$. In this section, an additional property of this contrast is proved. It will be shown that the set of local maximum points of $f(\mathbf{w})$ subject to $\mathbf{w} \in \mathcal{V}_K^1$ *coincides* with $\{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$. Since the proof is quite involved, the methodology is first sketched in Section IV-E1. Then, the detailed results will be given in Section IV-E2.

1) Methodology: To prove some results of Section IV-E2, we will compute $\Delta f(\mathbf{w} + \delta\mathbf{w}, \mathbf{w})$ where $\mathbf{w}, \mathbf{w} + \delta\mathbf{w}$ are "close" vectors, i.e., where $\delta\mathbf{w}$ is an infinitesimal vector in the sense that $\|\delta\mathbf{w}\| > 0$ can be chosen as close as possible to zero.

More precisely, we will focus on $\mathbf{w} \in \mathcal{V}_K^1$ and we restrict $\delta\mathbf{w}$ to be of the form

$$\delta\mathbf{w}_{ij}^{\zeta,} \triangleq \zeta\mathbf{e}_i + \xi\mathbf{e}_j \tag{12}$$

for two given distincts $1 \leq i$ and $j \leq K$. In (12), ζ and ξ denote infinitesimal scalar numbers, satisfying $\mathbf{w} + \delta\mathbf{w}_{ij}^{\zeta} \in \mathcal{V}_K^1$. It is shown in Lemma 1 that for all $\mathbf{w} \in \mathcal{V}_K^1$, all distinct indexes i, j and sufficiently small $|\zeta| > 0$, then such $\delta\mathbf{w}_{ij}^{\zeta}$ can be found, yielding ξ and $\Delta f(\mathbf{w} + \delta\mathbf{w}_{ij}^{\zeta}, \mathbf{w})$. Next, Theorem 3 shows that for all $\mathbf{w} \in \mathcal{V}_K^1$, it always exists $\delta\mathbf{w}_{ij}^{\zeta}$, such that $\Delta f(\mathbf{w} + \delta\mathbf{w}_{ij}^{\zeta}, \mathbf{w}) > 0$ if $|\zeta| > 0$, provided that $\mathbf{w} \notin \{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$. Finally, considering several established results, Corollary 2 states that $\mathcal{A}_4$ holds.

2) Detailed Results:

Lemma 1: For all vectors $\mathbf{w} \in \mathcal{V}_K^1$ and two distinct indexes $1 \leq i$ and $j \leq K$, it exists two infinitesimal scalar numbers ζ and ξ such that for all $\epsilon > 0$, $\|\delta\mathbf{w}_{ij}^{\zeta}\| < \epsilon$ and $\mathbf{w} + \delta\mathbf{w}_{ij}^{\zeta} \in \mathcal{V}_K^1$.

For a given infinitesimal ζ

$$\xi = -\mathbf{w}(j) + \sqrt{\mathbf{w}(j)^2 - (2\mathbf{w}(i)\zeta + \zeta^2)} \tag{13}$$

and

$$\Delta f(\mathbf{w} + \delta\mathbf{w}_{ij}^{\zeta}, \mathbf{w}) = R(S_i)\zeta + R(S_j)\xi. \tag{14}$$

The proof is straightforward and is given in Appendix III.

Theorem 3: For all $\mathbf{w} \in \mathcal{V}_K^1 \setminus \{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$, there exist two distinct indexes $1 \leq i$ and $j \leq K$ such that $0 < \mathbf{w}(i), \mathbf{w}(j) < 1$. For such indexes, consider the infinitesimal vectors $\delta\mathbf{w}_1$ and $\delta\mathbf{w}_2$ defined as

$$\delta\mathbf{w}_1 \triangleq \delta\mathbf{w}_{ij}^{\zeta}$$
$$\delta\mathbf{w}_2 \triangleq \delta\mathbf{w}_{ij}^{-\zeta}$$

where $\delta\mathbf{w}_{ij}^{\zeta}(j)$ is given by ξ in (13) and $\delta\mathbf{w}_{ij}^{-\zeta}(j)$ is given by the same equation with ζ replaced by $-\zeta$. By Lemma 1, $\{\mathbf{w} + \delta\mathbf{w}_1, \mathbf{w} + \delta\mathbf{w}_2\} \subset \mathcal{V}_K^1$. The associated contrast variations are noted

$$\Delta f^1 \triangleq \Delta f(\mathbf{w} + \delta\mathbf{w}_1, \mathbf{w})$$
$$\Delta f^2 \triangleq \Delta f(\mathbf{w} + \delta\mathbf{w}_2, \mathbf{w}).$$

Then, if $\zeta > 0$, either $\Delta f^1 > 0$ or $\Delta f^2 > 0$.

The proof is relegated in Appendix IV.

Corollary 2 (Discriminant Contrast Property): Function f is a discriminant contrast in the sense that, under the whitening constraint, $Y_i \propto S_j$ if and only if $\mathbf{b}_i$ locally maximizes $\mathcal{C}(\mathbf{b}_i)$. By Theorem 1, it is known that the global maximum of $f(\mathbf{w})$

is reached for $\mathbf{w} \in \{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$, where k is defined in $\mathcal{A}_0$. Theorem 2 indicates that $f(\mathbf{w})$ admits a local maximum when $\mathbf{w} \in \{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$. Therefore, using $\mathcal{A}_1$, a local maximum of $f(\mathbf{w})$ exists when $Y_i \propto S_j$. Finally, by Theorem 3, no local maximum exists for $\mathbf{w} \notin \{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$ (i.e., when $Y_i \neq \lambda S_j$, $\lambda \in \mathbb{R}_0$), which proves the corollary.

According to the discriminacy property, if $\mathcal{C}(\mathbf{b}_i) = f(\mathbf{c}_i)$ is maximized locally (e.g., by gradient–ascent), then the ith output Y_i must be proportional to one of the S_j.

Remark 2: (Restriction of $\mathbf{B} \in \mathbb{R}^{K\times K}$ to $\mathbf{B} \in \mathcal{O}(K)$) When proving the previous results, it is not always constrained that the $\mathbf{c}_i$ must satisfy another condition than $\mathbf{c}_i \in \mathcal{V}_K^\lambda$. However, in order to avoid extracting twice the same source, the $\mathbf{c}_i$ can be always kept orthonormal: we could, e.g., constrain $\mathbf{B}$ to belong either to $\mathcal{O}(K)$ or to $\mathcal{SO}(K)$. Hence, a natural question arises: Do $\mathcal{A}_1 - \mathcal{A}_4$ still hold under the additional constraint that $\mathbf{B}$ must belong to $\mathcal{O}(K)$ or to $\mathcal{SO}(K)$? Clearly, $\mathcal{A}_1$ is fulfilled, as well as $\mathcal{A}_2$, since the global maximum point is also included in $\mathcal{SO}(K) \subset \mathcal{O}(K) \subset \mathbb{R}^{K\times K}$. This can be extended to the local maximum points $\mathbf{w} \in \{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$ of $f(\mathbf{w})$ if $\mathbf{w} \in \mathcal{V}_K^1$. Indeed, since a manifold is a topological space which is locally Euclidean, for all $\mathbf{B} \in \mathcal{O}(K)$, the restriction of the neighborhood of $\mathbf{B}$ to the manifold induced by $\mathcal{O}(K)$ is a subset of the neighborhood of $\mathcal{B}$ in the whole $\mathbb{R}^{K\times K}$ space (recall that $\mathcal{O}(K) \subset \mathbb{R}^{K\times K}$). This is also true for $\mathbf{B} \in \mathcal{SO}(K)$, since $\mathcal{SO}(K)$ is a connected subgroup of $\mathcal{O}(K)$ and because it is a Lie group [33]; hence, it is also a smooth manifold [34].[2]

The only result that still has to be proved is $\mathcal{A}_4$, i.e., no local maximum point exists on the contrast restricted to $\mathcal{O}(K)$ for $\mathbf{w} \in \mathcal{V}_K^1/\{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$. To prove that, we assume that $p-1$ sources have been recovered, and, thus, we consider a matrix $\mathbf{B} \in \mathcal{O}(K)$ which is arbitrary except that, without loss of generality, its $p-1$ first rows correspond to the extraction of the first $p-1$ sources and that the pth output is not yet a source: $\mathbf{c}_p \in \mathcal{V}_K^1/\{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$. We will prove that, for such a matrix, it always exists a direction belonging to $\mathcal{O}(K)$ such that the contrast can be increased if $\mathbf{B}$ is updated by an infinitesimal way in that direction. Clearly, because of the orthogonality constraint of $\mathbf{B}$ (and thus of $\mathbf{C}$) and the extraction of the first $p-1$ sources, by hypothesis, the first $p-1$ entries of $\mathbf{c}_p$ are zero and must remain so. Any update that modifies only the $K-p+1$ last entries of $\mathbf{c}_p$ will thus preserves the orthogonality constraint, up to an orthogonalization of the $K-p$ last rows of $\mathbf{B}$. In particular, by taking $i, j \in \{p, \ldots, K\}$, the updates $\mathbf{c}_p \leftarrow \mathbf{c}_p + \delta\mathbf{w}_1$ or $\mathbf{c}_p \leftarrow \mathbf{c}_p + \delta\mathbf{w}_2$ considered in Theorem 3 satisfy the unit-norm $\|\mathbf{c}_p\| = 1$ contraint (by Lemma 1) and the orthogonality with the previous rows: $\mathbf{c}_p\mathbf{c}_q^T = \delta_{p,q}$, for all $1 \leq p < q$. Hence, one can always perform these updates without violating the constraints, and Theorem 3 ensures that the contrast function is increased in at least one of the analyzed situations.

Therefore, all the properties of $\mathcal{C}(\mathbf{b}_i)$ analyzed in $\mathbb{R}^{K\times K}$ subject to $\mathbf{b}_i \in \mathcal{V}_K^1$ still hold when one restricts $\mathbf{B}$ to be in the lower dimensional subset $\mathcal{O}(K) \in \mathbb{R}^{K\times K}$ or $\mathcal{SO}(K) \in \mathbb{R}^{K\times K}$. The last restrictions avoid to extract twice the same source.

[2]Note that we can talk about the restriction to $\mathcal{O}(K)$ of a neighborhood of $\mathbf{B} \in \mathcal{O}(K)$. Indeed, even if $\mathcal{O}(K)$ is a *disconnected* group, the last is built from two *connected* components: the special orthogonal group $\mathcal{SO}(K)$ $(\det \mathbf{B} = 1)$ including rotation matrices and the set $\mathcal{O}(K) \setminus \mathcal{SO}(K)$ including improper rotation matrices $(\det \mathbf{B} = -1)$.

V. Practical Issues

In practice, the output ranges are unknown. Indeed, from (5), the contrast depends on the $R(S_i)$. Therefore, the range has to be computed from the sample set, and a careful and reliable estimation of R is necessary to guarantee that the estimated range will satisfy the properties of the exact range quantity. Range estimation (also called endpoint estimation problem) is known to be a difficult task; it has been extensively studied in statistics and econometrics. However, most of the proposed methods require resampling or other computationally intensive techniques, involving tricky tuning of parameters. Moreover, specific assumptions on the density tails are usually needed. Such estimators do not really match the ICA requirements, since they are quite slow and nonblind. In addition, in the ICA context, it must be stressed that the output pdf (i.e., the one for which we have to estimate the range) varies due to the iterative updates of the demixing matrix rows $\mathbf{b}_i$.

In this paper, we will focus on range estimation approaches using order statistics, even though it is possible to consider other kinds of estimators. The simplest way for estimating the range of a random variable X based on a finite sequence of observations of size N: $\mathcal{X}_N = \{x_1, x_2, \cdots, x_N\}$ is to compute the observed range, that is $R^\star(X)$.

This statistical quantity can be rewritten using the order statistic notations. Let us suppose that $\mathcal{X}_{(N)}$ is an ordered version of $\mathcal{X}_N$, where the elements $x_{(i)}$ are ordered by increasing values, that is $\mathcal{X}_{(N)} = \{x_{(1)}, x_{(2)}, \cdots, x_{(N)}\}$ with $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(N)}$. Then, we have

$$R^\star(X) \triangleq x_{(N)} - x_{(1)}.$$

The main problem of this estimator is that it is highly sensitive to noise and outliers. Even though the samples are not set to the fourth power as when dealing with the kurtosis, there is no regularization induced by the interior points: only the two observed extreme values are taken into account. Therefore, another estimator can be used

$$R_m^\star(X) \triangleq x_{(N-m+1)} - x_{(m)}$$

where $m < \lfloor N/2 \rfloor$, $m \in \mathbb{Z}$. However, similarly to $R_1^\star(X) = R^\star(X)$, $R_m^\star(X)$ is based only on two sample points, and has obviously a higher bias than $R^\star(X)$. Hence, we suggest the following estimator using $2m$ sample points:

$$\langle R_m^\star(X) \rangle \triangleq 1/m \sum_{i=1}^{m} R_i^\star(X)$$

with this estimator of the range, the finite sample approximation of contrast $\hat{\mathcal{C}}(\mathbf{b}_i)$ becomes

$$\hat{\mathcal{C}}^\star(\mathbf{b}_i) = -\langle R_m^\star(\mathbf{b}_i\mathbf{X}) \rangle \quad \text{s.t. } \mathrm{Var}(\mathbf{b}_i\mathbf{X}) = 1. \tag{15}$$

In order to study the behavior of $\hat{\mathcal{C}}^\star$, let us analyze some properties of the $\langle R_m^\star(\mathbf{b}_i\mathbf{X}) \rangle$ estimator.

A. Some Properties of $\langle R_m^\star(X) \rangle$

To analyze the theoretical behavior of $R_m^\star$, we should consider $x_{(i)}$ as a realization of the random variable $X_{(i)}$ with density $f_{X_{(i)}}(r)$. Hence, we should deal with $R_i^\star(X) = X_{(N-i+1)} - X_{(i)}$. The density $f_{R_m^\star(X)}(r)$ can be

TABLE I
FIVE UNIT-VARIANCE RANDOM VARIABLES: THEIR PDF, SUPPORT, AND RANGE (NOTE: $f_X(\zeta) = 0$ IF $\zeta \notin \Omega(X)$). THE PDF AND CDF ARE PLOTTED IN FIG. 3 WHERE THE SUPPORTS ARE SCALED TO BE INCLUDED IN (0,1)

Density	Notation	Analytical form of the pdf	Support $\Omega(.)$	Range $R(.)$
Uniform	$U \sim f_U$	$f_U(\zeta) = \frac{1}{2\sqrt{3}}$	$[-\sqrt{3}, \sqrt{3}]$	$2\sqrt{3}$
Linear	$L \sim f_L$	$f_L(\zeta) = \frac{\zeta+\sqrt{8}}{9}$	$[-\sqrt{8}, \sqrt{8}/2]$	$3/2\sqrt{8}$
Triangular	$T \sim f_T$	$f_T(\zeta) = \frac{\zeta+\sqrt{6}}{6}$ if $\zeta \in <0$, $\frac{\sqrt{6}-\zeta}{6}$ if $\zeta > 0$	$[-\sqrt{6}, \sqrt{6}]$	$2\sqrt{6}$
"V"-shape	$V \sim f_V$	$f_V(\zeta) = -\zeta/2$ if $\zeta < 0$, $\zeta/2$ if $\zeta > 0$	$[-\sqrt{2}, \sqrt{2}]$	$2\sqrt{2}$
Bi-Uniform	$BU \sim f_{BU}$	$f_{BU}(\zeta) = 1/\alpha$	$[-\alpha, -\alpha/2] \cup [\alpha/2, \alpha]$, $\alpha = \sqrt{12/7}$	2α

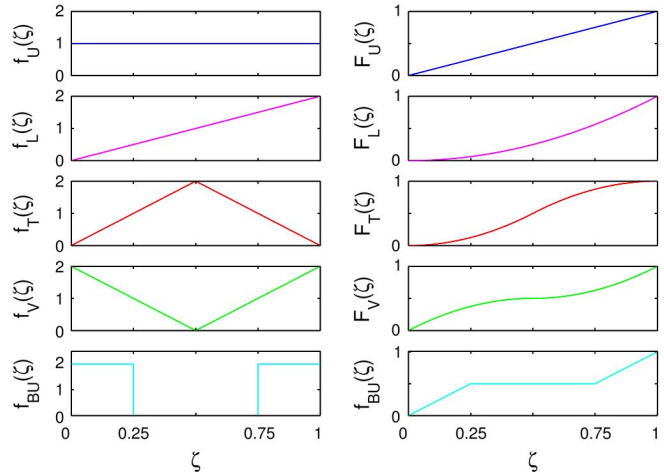

Fig. 3. Densities and cdf of five sources with equal support (here: [0,1]).

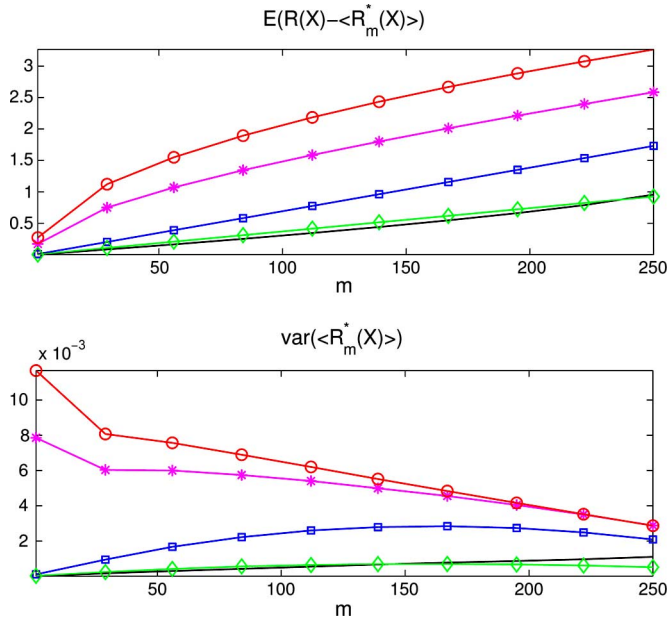

Fig. 4. Empirical expectations of the range estimation error and the variances of $\langle R^{\star}_m(X)\rangle$ for $N = 500$ and 1000 trials for the variables given in Table I scaled to unit variance: U (□), T (o), L (∗), V (no marker) and BU (⋄).

found through the order statistics density $f_{X_{(i)}}(r)$ computed from the parent pdf and cumulative density function (cdf).

The density of $\langle R^{\star}_m(X)\rangle$ could then be obtained by integrating the joint density of the quasi-ranges with *ad hoc* integration limits. However, this approach is of low use in practice since most often, no analytical expression can be found for $f_{\langle R^{\star}_m(X)\rangle}$.

A simple way to circumvent this problem is to use numerical simulations and to work with the $x_{(i)}$, the ith largest realization of X from a sample of size N, rather than with the random variable $X_{(i)}$ that depends on f_X. Let us consider five unit-variance random variables U, L, T, V, and BU having, respectively, uniform (f_U), linear (f_L), triangular (f_T), V-shape (f_V), and "biuniform" (f_{BU}) densities (see Table I; the pdf, with support convex hulls mapped to [0,1] here for ease of readability, are plotted in Fig. 3).

The empirical expectations of the error of the range estimator and the variance of the estimator are plotted for the aforementioned random variables in Fig. 4. Observe that the lower the m is, the lower the error $R(X) - \langle R^{\star}_m(X)\rangle$ since the last criterion is positive and because $\Pr[\langle R^{\star}_m(X)\rangle \geq \langle R^{\star}_{m^{\star}}(X)\rangle] = 1$ for all $m \leq m^{\star}$. The error rate increases with m at a rate depending on the density. Though not visible in Fig. 4, it can reasonably be understood that for fixed m, $\langle R^{\star}_m(X)\rangle$ is an asymptotically unbiased estimator of $R(X)$ with increasing N, whatever is the distribution of X, provided that the extreme values of the support are not isolated points (the probability to observe a point in the neighborhood of the extreme points must be nonzero). However, the convergence rate depends on f_X.

B. On the Choice of m

In practice, the value of "m given N" for $\langle R^{\star}_m(.)\rangle$ estimator has to be carefully chosen. On one hand, for increasing N, there is no need to take a large value for m since the sample points tend to *fill* the distribution range (this is the methodology proposed by Devroye and Wise [35]). On the other hand, even if the estimation error increases with m, a large m could help canceling a possible additive noise effect. As shown in Fig. 4, in such a case, the range of a unit-variance triangular variable T estimated using $\langle R^{\star}_m(T)\rangle$ can be lower than the one of a uniform variable T, i.e., $\langle R^{\star}_m(T)\rangle \leq \langle R^{\star}_m(U)\rangle$, even if we obviously have $R(T) > R(U)$. This could be a crucial problem in the ICA application. Indeed, assume $K = 2$ and both S_1 and S_2 are uniform sources. Then, there exists a threshold $m^{\dagger}$ such that by choosing $m > m^{\dagger}$ (from simulation results, $m^{\dagger} \approx 100$ was observed for $N = 500$), the global maximum of $\hat{\mathcal{C}}^{\star}(\mathbf{b}_i)$ will be found for $\mathbf{b}_i^{\dagger}$ such that $\mathbf{b}_i^{\dagger}\mathbf{A} = \sqrt{2}/2[1,1]$ since $f_{\mathbf{b}_i^{\dagger}\mathbf{X}}$ is a unit-variance triangular signal. In other words, the sources are not recovered at all. This is a consequence of the following facts: 1) the range of variables with smoothly decreasing tails are much more difficult to estimate than densities taking high values near the boundaries and 2) the pdf of summed variables

is the convolution of the densities of the added variables, so that the tails of the output pdf tends to be less sharp than the source tails.

Therefore, we need some guidelines for choosing the largest possible value for m (for regularization purposes) but limiting the error on the range estimation by a threshold $\mathcal{E}$, at least in probability. The following empirical law is proposed for selecting a default value for m (see Appendix V, for more details):

$$m^{\sharp}(N) = \max\left(1, \left\lceil \Re\left\{\left(\frac{N-18}{6.5}\right)^{0.65}\right\} - 4.5 \right\rceil\right) \quad (16)$$

where $\overline{\alpha}$ denotes the nearest integer to α.

When N is large enough, and if m is not too large, the range estimation is reliable and $\hat{\mathcal{C}}^{\star}(\mathbf{b}_i)$ is close from $\mathcal{C}(\mathbf{b}_i)$. In this favorable context, both criteria share the same behavior with respect to the transfer matrix, including the discriminant contrast function property.

VI. Maximization of the Contrast

The $\mathcal{C}(\mathbf{b}_i)$ contrast is not everywhere differentiable, due to the absolute values. Hence, gradient-based ICA algorithms cannot be used for maximizing it; the desired solutions are not stationary points of these algorithms. On the other hand, since we can focus on unmixing matrices $\mathbf{B} \in \mathcal{O}(K)$, one can proceed to a geodesic optimization of $\mathcal{C}(\mathbf{B})$ on the Stiefel manifold. Because of the Lie group structure of $\mathcal{O}(K)$ [33], for any pair of matrices $\mathbf{B}$ and $\mathbf{G}$ in $\mathcal{O}(K)$, $\mathbf{GB} \in \mathcal{O}(K)$. Therefore, a geodesic optimization can be obtained by factorizing $\mathbf{B}$ as a product of Givens rotation matrices $\mathbf{G}_{ij}^{\alpha}$ $(i < j)$ and by updating the angle α according to $\mathcal{C}$

$$\mathbf{B} \leftarrow \mathbf{G}_{ij}^{\alpha}\mathbf{B}.$$

Recall that the Givens matrix is a rotation matrix equal to the identity except entries $[\mathbf{G}_{ij}^{\alpha}]_{ii} = [\mathbf{G}_{ij}^{\alpha}]_{jj} = \cos(\alpha)$ and $[\mathbf{G}_{ij}^{\alpha}]_{ij} = -[\mathbf{G}_{ij}^{\alpha}]_{ji} = \sin(\alpha)$. With such matrices $\mathbf{G}_{ij}^{\alpha}$ and if the initial value of $\mathbf{B}$ is in $\mathcal{SO}(K)$, then, at each step, $\mathbf{B}$ belongs to the connected subgroup $\mathcal{SO}(K)$ of $\mathcal{O}(K)$ [33].

A lot of different methods for maximizing our nondifferentiable contrast on $\mathcal{SO}(K)$ have been tried, using, among others, discrete-gradient approximations based on a second-order Taylor expansion. Unfortunately, they lead to disappointing results, mainly because of the difficulty to obtain a good estimate of the derivative of the contrast function. Moreover, these algorithms involve several additional parameters that are tedious to adjust, such as the finite difference in the computation of the discrete derivative. On the contrary, a very simple algorithm, first sketched in [36], gave the best separation results. It is recalled in Section VI-A and its performances on bounded sources are presented in Section VI-C.

A. Algorithm

The algorithm assumes that the observed mixtures $\mathbf{X}$ are whitened, and proceeds to a contrast maximization by always keeping $\mathbf{B} \in \mathcal{SO}(K)$. The proposed algorithm is able to maximize any continuous but not necessarily differentiable componentwise contrast. In the present case, we focus on $\hat{\mathcal{C}}^{\star}(\mathbf{b}_i)$ given by (15).

ICAforNDC($\mathcal{C}, \mathbf{X}, \beta, \tau$)
Input: $\mathcal{C}$ (contrast function),
$\mathbf{X}$ (whitened mixtures),
β (convergence parameter, default: 0.75),
τ (number of iterations, default: 50).
Output: $\mathbf{B}$ (separation matrix).
Auxiliary: α (an angle),
K (number of sources), i, j, t (iteration indices),
$\mathbf{G}_{ij}^{\alpha}$ (a Givens matrix).
Begin
▷ Initialize $\mathbf{B}$ to the identity matrix.
$\mathbf{B} \leftarrow \mathbf{I}_K$
▷ Deflation approach: estimate each source sequentially.
for $i \leftarrow 1$ **to** K **do**
  ▷ Iterate for the i-th source
  for $t \leftarrow 1$ **to** τ **do**
    ▷ Set the current angle variation.
    $\alpha \leftarrow \pi\beta^t$
    ▷ Look in each direction.
    for $j \leftarrow i+1$ **to** K **do**
      ▷ Determine best contrast value
      if $\mathcal{C}(\mathbf{G}_{ij}^{+\alpha}\mathbf{B}) > \mathcal{C}(\mathbf{B})$ & $\mathcal{C}(\mathbf{G}_{ij}^{+\alpha}\mathbf{B}) > \mathcal{C}(\mathbf{G}_{ij}^{-\alpha}\mathbf{B})$ **then**
        ▷ Update the i-th and j-th rows of $\mathbf{W}$.
        $\mathbf{B} \leftarrow \mathbf{G}_{ij}^{+\alpha}\mathbf{B}$
      else if $\mathcal{C}(\mathbf{G}_{ij}^{-\alpha}\mathbf{B}) > \mathcal{C}(\mathbf{B})$ & $\mathcal{C}(\mathbf{G}_{ij}^{-\alpha}\mathbf{B}) > \mathcal{C}(\mathbf{G}_{ij}^{+\alpha}\mathbf{B})$ **then**
        ▷ Update the i-th and j-th rows of $\mathbf{W}$.
        $\mathbf{B} \leftarrow \mathbf{G}_{ij}^{-\alpha}\mathbf{B}$
      end if
    end for
  end for
end for
Return $\mathbf{B}$
End

Fig. 5. Pseudocode for the deflation ICA algorithm for nondifferentiable contrast functions (comments begin with a triangle) [36]. The mixture vector $\mathbf{X}$ is used in the evaluation of $\mathcal{C}$.

The maximization procedure considers the unmixing matrix as a set of orthonormal vectors and is based on pairwise angular variations of these vectors (Jacobi-like rotations). In order to remain meaningful, the optimization procedure of the contrast function relies on the two following assumptions. First, the contrast function should be continuous. Second, it should also be discriminant. On the other hand, it is not assumed that the contrast function is differentiable with respect to α. Therefore, the contrast function may be a piecewise linear function (discontinuous derivative), just like $\hat{\mathcal{C}}^{\star}(\mathbf{b}_i)$.

Under the aforementioned assumptions, the simple algorithm in Fig. 5 may be used to compute each row of $\mathbf{B}$. As it can be seen, the algorithm keeps $\mathbf{B}$ orthogonal. The only parameters of the algorithm are τ and β, which are, respectively, the number of iterations and an exponentially decaying learning rate. Usually, with the default values given in Fig. 5, the algorithm has converged after ten or 20 iterations $(10 < \tau < 20)$. By construction, the algorithm is monotonic: the contrast is either decreased or kept constant.

B. Discriminacy and Jacobi Updates

In Remark 2, it is explained that under the $\mathbf{B} \in \mathcal{O}(K)$ or $\mathbf{B} \in \mathcal{SO}(K)$ constraint, at least one direction always exists that allows one to increase the contrast function, provided that $\mathbf{w} \notin \{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$. However, even if Jacobi updates may yield any $\mathbf{B} \in \mathcal{SO}(K)$ if $\mathbf{B}$ is initialized to a point in $\mathcal{SO}(K)$, as, e.g., $\mathbf{I}_K$, the corresponding trajectories are not arbitrary on the associated

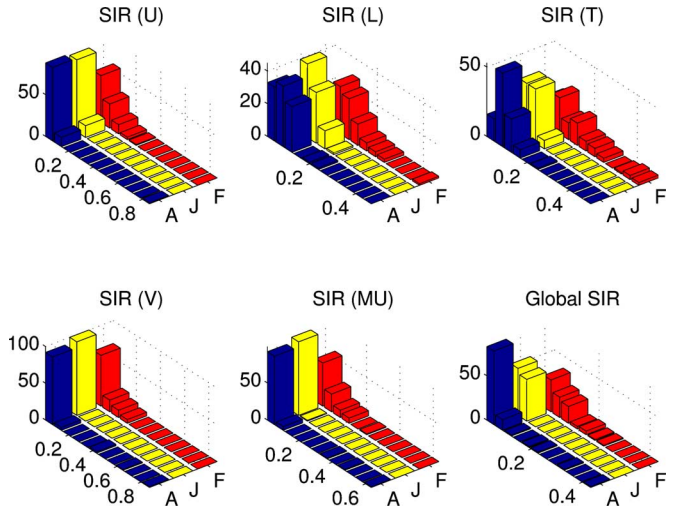

Fig. 6. 12-bins histograms of PI for each extracted source, for 100 trials, $N = 2000$, and $m = m^\sharp(N) = 37$. The analyzed algorithms are AVOSICA (A), JADE (J), and FastICA (F). The *global PI* is the averaged PI computed from the individual source PIs for a given trial.

manifold: only updates of the form $\mathbf{c}_p \rightarrow \cos\alpha\mathbf{c}_p + \sin\alpha\mathbf{c}_q$ are made possible. Therefore, along Jacobi trajectories, a contrast can seem to have a local maximum at a given point, even if this contrast can be increased along another trajectory of the restriction of $\mathcal{O}(K)$ to $\mathcal{SO}(K)$ (just think about a toy contrast including a saddle point). However, the probability to get stuck in such a false local maximum point is very low, except for "pathological" contrast functions: many directions are already explored using Jacobi rotations. This hypothetical problem is just emphasized for the sake of completeness.

C. Simulations

Several range estimators and optimization algorithms have been tried; AVeraged Order Statistics ICA (AVOSICA), which is the optimization scheme obtained from the combination of the contrast $\hat{\mathcal{C}}^\star(\mathbf{b}_i)$ with the algorithm of Fig. 5, gave the best performances. Therefore, we shall compare AVOSICA to JADE and FastICA (with Gaussian nonlinearity, because of its robustness [1] and because it gave the best results among all the available nonlinearities in the FastICA package).

The algorithms have been tested on the extraction of five bounded and white sources (with different densities) from five mixtures. The pdf and cdf of these sources (mapped to the [0,1] interval) are illustrated in Fig. 3. The mixing matrix is built from 25 coefficients drawn from a uniform distribution on (0,1).

Fig. 6 compares the histograms of the performance index (PI) for each extracted source in the noise-free case for $N = 2000$ and $m = m^\sharp(N)$; the lower is the PI, the better is the separation. Recall that after having solved the permutation indetermination (which is possible on toy examples), the PI criterion of the ith source reduces to $\mathrm{PI}(S_i) = \sum_{j\neq i} |[\mathbf{C}]_{ij}|/|[\mathbf{C}]_{ii}|$. Clearly, the lower is the highest bin coordinate, the better is the method in average, and a narrow spectrum indicates a low variance among the obtained results. Then, for an ideal separation method, a single bin of height equal to the number of trials, located close to zero, would be observed. We can observe in Fig. 6 that AVOSICA gives the most interesting results, in comparison to JADE and FastICA, especially for the separation of sources with linear and triangular pdf. Table II summarizes the average global PI of ICA algorithms for various noise levels. Since we deal with PI, the performance results are analyzed from the mixing matrix recovery point of view; the source denoising task is not considered here. The global PI, for a given trial, is obtained by computing the mean of the extracted sources PI. The good results of AVOSICA can be observed, despite the fact that the value of m has not been chosen to optimize the results, i.e., we always have taken $m = m^\sharp(N)$ given by (40). It must be stressed that the value of the parameter m is not critical when chosen around $m^\sharp(N)$. JADE is a very good alternative when the dimensionality of the source space is low. The computational time of FastICA is its main advantage.

TABLE II
100-TRIALS EMPIRICAL MEANS AND VARIANCES OF GLOBAL PI OF SEVERAL ICA ALGORITHMS (*GLOBAL PI* IS THE AVERAGED PI COMPUTED FROM THE INDIVIDUAL SOURCE PIS FOR A GIVEN TRIAL); $m = m^\sharp(N)$. GAUSSIAN NOISE WITH STANDARD DEVIATION σ_n HAS BEEN ADDED TO THE WHITENED MIXTURES (SO THAT FOR A GIVEN σ_n, THE MIXTURE SNRS EQUAL $-10\log\sigma_n^2$; THEY DO NOT VARY BETWEEN TRIALS, AND DO NOT DEPEND ON THE MIXING WEIGHTS). THE NUMBERS BETWEEN PARENTHESES REFLECT THE VARIANCE; THE PERFORMANCES OF THE WINNER ALGORITHM ARE IN BOLDFACE

σ_n^2	N	AVOSICA	JADE	FastICA
0	500	**.107 (.004)**	.13 (.005)	.195 (.015)
	2000	**.049** (.004)	.056(**.0005**)	.104 (.005)
0.01	500	**.11 (.003)**	.13 (.005)	.19 (.02)
	2000	**.05** (.001)	.06 (**.0006**)	.105 (.006)
0.05	500	**.108 (.0027)**	.122 (.0032)	.195 (.019)
	2000	**.051 (.0006)**	.06 (**.0006**)	.112 (.004)

Remark 3 (Complexity of the Algorithms): Depending on how the q lowest and highest out values are computed, the complexity of AVOSICA is either $O(K^2\tau(KN + N\log N))$ with a complete sort operation or $O(K^2\tau(KN + q\log N))$ with a partial sort. In these complexities, τ is the number of iterations in each of the K deflation stages. Computing the output $\mathbf{Y}_i$, needed before each update, requires KN operations. By comparison, the FastICA algorithm has a complexity of $O(K\tau(KN+N+K^2))$ (the three terms in the rightmost factor correspond, respectively, to the computation of the output, of the kurtosis, and the Gram–Schmidt orthonormalization). Of course, as FastICA involves a fixed-point optimization scheme, τ is usually much lower than in AVOSICA. Finally, JADE has a complexity of $O(K^4N + \tau K^5)$, where the two terms correspond, respectively, to the computation of all cross cumulants and to the approximate joint diagonalization.

VII. DISCUSSION

Section VI emphasizes the good performances of the method for the noise-free and low-noise bounded source separation cases. In spite of these interesting results, we have to mention that range estimation techniques have poor performances when the finite sample sequence has few points near the scatter plot corners. For instance, some problems could be encountered in practice when dealing with bounded sources for which only small samples are available if the source densities have long flat tails. If we suppose $K = 2$ and if we look at the joint

scatter plot of sources sharing the same density with long flat tails, a four-branches star would be observed, with axes colinear to the source axes. However, in spite of the theoretical results obtained from $\mathcal{C}(\mathbf{b}_i)$, the axes along which the estimated range width projection $\hat{\mathcal{C}}^{\star}(\mathbf{b}_i)$ is minimum are no more the source axes ($\theta = k\pi/2$) but rather the diagonal directions ($\theta = (2k+1)\pi/4$), so that the method totally fails [37]. The problem is due to the fact that few points are observed into the corner of the joint pdf. For such kind of sources, the proposed method is not really adapted, because more sophisticated range estimators are needed (other techniques, such as, e.g., the one proposed in [20], have to be preferred in this specific case).

Finally, in addition to the separation performance improvement when dealing with bounded sources, the proposed approach has three major advantages. First, the method can be extended to separate correlated signals, provided that some sample points can be observed in the corners of the joint scatter plot [37]. For instance, two correlated images can be separated with largely higher separation performances than when using usual ICA algorithms. Note that the unmixing matrix has to be postprocessed because the source uncorrelation assumption is not valid here; the orthogonality constraint between the rows of the unmixing matrix can be relaxed (for more details about this, we refer to [37]). Second, when the sources densities are strongly bimodal, it is known that usual ICA algorithms based on the minimum mutual information or maximum negentropy approaches lead to spurious solution [17], [41]. The proposed method is proved to be free of spurious maxima, as shown by the *discriminant contrast property*. Third, it should be stressed that the method is very robust to the dimensionality of the source space; a variant of AVOSICA has been tested on the MLSP 2006 competition benchmark (evaluation of ICA algorithms for large-scale, ill-conditioned, and noisy mixtures). It has outperformed the results of all other algorithms that were tested in the competition (see [38], for more details).

VIII. Conclusion

In this paper, a new objective function for source separation is proposed, based on the output ranges. The contrast properties of the criterion have been proved, and the discriminacy property ensures that no spurious solution can be obtained, provided that the range width is estimated in a satisfactory way. The method is related to the state of the art and some relationships with other well-known approaches to ICA have been drawn. In practice, the range estimation is a difficult task, but it is shown that a simple batch algorithm based on averaged order-statistic differences can be used for the separation of various kinds of sources (i.e., with various bounded densities). A default value has been proposed for the number of order statistics that has to be taken into account when robustness is needed.

Appendix I
Proof of Theorem 1

The proof of Theorem 1 will be based on Propositions 1 and 2, and assumes $R(S_1) = R(S_2) = \ldots = R(S_k)$ with $R(S_k) < R(S_{k+1})$.

The first proposition will show that if a vector $\mathbf{w}$ has a nonzero entry at any place $i > k$, then $\mathbf{w}$ does not correspond to a global maximum point of f. The second proposition shows that among all the remaining vectors candidate to be a global maximum point (thus satisfying necessarily $\mathbf{w}(r) = 0$ for $k < r \leq K$), none can be a global maximum provided that it is proportional to the basis vector $\mathbf{e}_1, \ldots \mathbf{e}_k$ (the absolute value of the coefficient is given by the norm constraint).

Proposition 1: Let us define a $\mathbf{p} \in \mathcal{V}_K^{\lambda}$ vector respecting $\mathbf{p}(r) > 0$ for any $k < r \leq K$. Consider vector $\mathbf{q}$ defined by

$$\begin{cases} \mathbf{q}(r) = 0 \\ \mathbf{q}(k') = \sqrt{\mathbf{p}(k')^2 + \mathbf{p}(r)^2} \text{ with } 1 \leq k' \leq k \\ \mathbf{q}(j) = \mathbf{p}(j) \text{ for all } 1 \leq j \leq K, \qquad j \notin \{k', r\} \end{cases}. \quad (17)$$

Then, $\mathbf{q} \in \mathcal{V}_K^{\lambda}$ and $f(\mathbf{q}) > f(\mathbf{p})$: i.e., $\mathbf{p} \notin \{\mathbf{w} : \mathbf{w} = \arg\max_{\{\mathbf{w} \in \mathcal{V}_K^{\lambda}\}} f(\mathbf{w})\}$

Proof: It is trivial to show that $\mathbf{q} \in \mathcal{V}_K^{\lambda}$. On the other hand, we have $\mathbf{p}(r)^2 R^2(S_{k'}) < \mathbf{p}(r)^2 R^2(S_r)$ and

$$\begin{aligned} \left(\mathbf{p}(k')^2 + \mathbf{p}(r)^2\right) R^2(S_{k'}) < \mathbf{p}(k')^2 R^2(S_{k'}) + \mathbf{p}(r)^2 R(S_r) \\ + \underbrace{2\mathbf{p}(k')\mathbf{p}(r) R(S_{k'}) R(S_r)}_{\geq 0} \end{aligned}$$

$$R(S_{k'})\sqrt{\mathbf{p}(k')^2 + \mathbf{p}(r)^2} < \mathbf{p}(k') R(S_{k'}) + \mathbf{p}(r) R(S_r). \quad (18)$$

Hence, it results from the definition of $\mathbf{q}$ that $-f(\mathbf{q}) < -f(\mathbf{p})$ and thus $f(\mathbf{q}) > f(\mathbf{p})$. ■

Proposition 2: For any $\mathbf{p} \in \mathcal{V}_K^{\lambda}$ vector satisfying $\mathbf{p}(j) = 0$ for all $k < j \leq K$, then $f(\mathbf{p}) \leq f(\lambda \mathbf{e}_j)$, $1 \leq j \leq k$ with equality if and only if $\mathbf{p} \in \{\lambda.\mathbf{e}_1, \ldots, \lambda.\mathbf{e}_k\}$

Proof: If $\mathbf{p}(j) = 0$ for all $j > k$, then, because $\mathbf{p} \in \mathcal{V}_K^{\lambda}$, it must exist $r \leq k$ such that $\mathbf{p}(r) > 0$. On the other hand, for any $1 \leq r \neq r' \leq k$, we know that $\mathbf{p}(r') \geq 0$. Hence, by definition of k

$$\mathbf{p}(r) R(S_r) + \mathbf{p}(r') R(S_{r'}) = \left(\mathbf{p}(r) + \mathbf{p}(r')\right) R(S_r). \quad (19)$$

Let us define $\mathbf{q}$ by $\mathbf{q}(j) = \mathbf{p}(j)$ for $j \notin \{r, r'\}$, $\mathbf{q}(r) = \sqrt{\mathbf{p}(r)^2 + \mathbf{p}(r')^2}$ and $\mathbf{q}(r') = 0$. Then, it is straightforward to show that $\mathbf{q} \in \mathcal{V}_K^{\lambda}$, and that $f(\mathbf{q}) \geq f(\mathbf{p})$ with equality if and only if $\mathbf{p}(r') = 0$. To prove the last claim, remark that

$$\sqrt{\mathbf{p}(r)^2 + \mathbf{p}(r')^2} \leq \mathbf{p}(r) + \mathbf{p}(r') \quad (20)$$

with equality only when $\mathbf{p}(r') = 0$. Hence, by iterating this result setting $\mathbf{p} \leftarrow \mathbf{q}$, if such a $\mathbf{p}$ vector has at least two strictly positive elements, then $f(\mathbf{p}) < f(\lambda \mathbf{e}_j)$, with $1 \leq j \leq k$. On the other hand, it is easy to see that if a $\mathbf{p}$ vector respecting $\mathbf{p}(k+1) = \ldots = \mathbf{p}(K) = 0$ and $\mathbf{p} \in \mathcal{V}_K^{\lambda}$ has a single nonzero entry, then $\mathbf{p} \in \{\lambda.\mathbf{e}_1, \ldots, \lambda.\mathbf{e}_k\}$. ■

By iterating Proposition 1, for any vector $\mathbf{p} \in \mathcal{V}_K^{\lambda}$ such that it exists $k < r \leq K$ with $\mathbf{p}(r) > 0$ it exists another vector $\mathbf{q} \in \mathcal{V}_K^{\lambda}$, respecting $\mathbf{q}(j) = 0$ for all $k < j \leq K$ satisfying $f(\mathbf{q}) > f(\mathbf{p})$. On the other hand, Proposition 2 shows that among all those $\mathbf{q}$ vectors, only $\mathbf{q} \in \{\lambda.\mathbf{e}_1, \ldots, \lambda.\mathbf{e}_k\}$ can maximize globally function f subjected to $\mathbf{q} \in \mathcal{V}_K^{\lambda}$.

Appendix II
Proof of Theorem 2

Proof: Suppose that $\hat{\mathbf{e}}_i \in \mathcal{V}_K^1$ is a vector close to $\mathbf{e}_i$, in the sense that $\hat{\mathbf{e}}_i = \mathbf{e}_i + \delta \mathbf{e}_i$ where $\delta \mathbf{e}_i$ is an infinitesimal vector. Obviously, $\delta \mathbf{e}_i(i) < 0$ and $\delta \mathbf{e}_i(j) \geq 0$, for $j \neq i$. We note

$\delta\mathbf{e}_i(i) = -\epsilon$ where $\epsilon > 0$ is an infinitesimal scalar. By the $||\hat{\mathbf{e}}_i|| = 1$ constraint, it shows that

$$\sum_{j\neq i}\hat{\mathbf{e}}_i(j)^2 = 1-\hat{\mathbf{e}}_i(i)^2 = 1-(1-\epsilon)^2. \tag{21}$$

On the other hand, by (11)

$$\Delta f(\mathbf{e}_i,\hat{\mathbf{e}}_i) = \underbrace{(1-\epsilon)R(S_i)+\sum_{j\neq i}\hat{\mathbf{e}}_i(j)R(S_j)}_{-f(\hat{\mathbf{e}}_i)} - \underbrace{R(S_i)}_{-f(\mathbf{e}_i)}. \tag{22}$$

Hence, Theorem 2 will be proven if

$$\sum_{j\neq i}\hat{\mathbf{e}}_i(j)R(S_j) > \epsilon R(S_i). \tag{23}$$

Let us denote the norm of $\hat{\mathbf{e}}_i$ subject to $j\neq i$ vector by

$$\lambda' \triangleq \sqrt{\sum_{j\neq i}\hat{\mathbf{e}}_i(j)^2}. \tag{24}$$

By (21), $\lambda' = \sqrt{1-(1-\epsilon)^2}$. Hence, by using Theorem 1 with $\mathbf{w} = [\hat{\mathbf{e}}_i(1),\ldots,\hat{\mathbf{e}}_i(i-1),\hat{\mathbf{e}}_i(i+1),\ldots,\hat{\mathbf{e}}_i(K)]$ and $\mathbf{w}\in\mathcal{V}^{\lambda'}_{K-1}$, then $f(\mathbf{w})\leq f(\lambda'\mathbf{e}_r)$, where $r=\arg\min_{j\neq i}\{R(S_j)\}$. In other words, the following inequality holds:

$$\sum_{j\neq i}\hat{\mathbf{e}}_i(j)R(S_j) \geq \underbrace{\sqrt{1-(1-\epsilon)^2}}_{\lambda'}R(S_r). \tag{25}$$

Then, having (23) in mind, a sufficient condition to prove Theorem 2 is to check that the following inequality holds for any sufficiently small $\epsilon>0$:

$$\lambda' R(S_r) > \epsilon R(S_i) \qquad \text{with } r\neq i. \tag{26}$$

By transitivity, the previous inequality holds when

$$\begin{aligned}\sqrt{2\epsilon-\epsilon^2}R(S_r) &> \epsilon R(S_i)\\ (2\epsilon-\epsilon^2)R^2(S_r) &> \epsilon^2 R^2(S_i).\end{aligned} \tag{27}$$

Hence, if $\epsilon[2R^2(S_r)-\epsilon(R^2(S_r)+R^2(S_i))]>0$ holds for any sufficiently small $\epsilon>0$, then (23) is fulfilled.

The last inequality is satisfied for all $0<\epsilon<(2R^2(S_r)/R^2(S_i)+R^2(S_r))$. This result concludes the proof: $\Delta f(\mathbf{e}_i,\hat{\mathbf{e}}_i)>0$ for all sufficiently small $\epsilon>0$. ■

Appendix III
Proof of Lemma 1

Proof: Let us fix the distinct indexes $1\leq i,j\leq K$ and the infinitesimal scalar ζ. Note that in some pathological cases, the sign of ζ cannot be arbitrarily chosen; otherwise, the $\mathbf{w}+\delta\mathbf{w}^{\zeta}_{ij}\in\mathcal{V}^1_K$ may be not satisfied (for example, if $\mathbf{w}=\mathbf{e}_i$, then we must obviously take $\zeta<0$ and $\xi>0$). The $\mathbf{w}+\delta\mathbf{w}^{\zeta}_{ij}\in\mathcal{V}^1_K$ constraint yields

$$\xi^2+2\mathbf{w}(j)\xi+\zeta^2+2\mathbf{w}(i)\zeta=0. \tag{28}$$

Both roots of (28) will lead to the same absolute value of $\mathbf{w}(r)+\delta\mathbf{w}^{\zeta}_{ij}(r)$, for all $1\leq r\leq K$. We focus on the single root of (28) satisfying $|\xi|<\mathbf{w}(j)$ ($\mathbf{w}+\delta\mathbf{w}^{\zeta}_{ij}\in\mathcal{V}^1_K$), which gives (13).

With this value of ξ, observe that $||\delta\mathbf{w}^{\zeta}_{ij}||\to 0$ as $|\zeta|\to 0$ (observe that there is no restriction to make ζ tending to zero since $\mathcal{V}^1_K$ is a connected set). This results from the fact that $\mathcal{V}^{\lambda}_K$ defines the surface of the K-dimensional sphere centered at the origin with radius λ in R^K_+, i.e., a continuous manifold in R^K_+.

Finally, by definition of $\delta\mathbf{w}^{\zeta}_{ij}$, the rth entry of $\mathbf{w}$ equals the rth entry of $\mathbf{w}+\delta\mathbf{w}^{\zeta}_{ij}$ except if $r\in\{i,j\}$, which gives the $\Delta f(\mathbf{w}+\delta\mathbf{w}^{\zeta}_{ij},\mathbf{w})$ given in the Lemma. ■

Appendix IV
Proof of Theorem 3

Proof: We freely assume $\zeta>0$. If $\Delta f^1>0$, the Theorem is obviously trivially proven. Consider then the unique alternative $\Delta f^1\leq 0$; we will show that in this case, $\Delta f^2>0$.

Combination of (13) and (14) with $\mathbf{w}\notin\{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_K\}$, $\Delta f^1\leq 0$, and ζ, a strictly positive infinitesimal scalar, give

$$\begin{aligned}-\Delta f^1 &= R(S_i)\zeta + R(S_j)\\ &\quad\times\left(-\mathbf{w}(j)+\sqrt{\mathbf{w}(j)^2-(2\mathbf{w}(i)\zeta+\zeta^2)}\right)\geq 0.\end{aligned} \tag{29}$$

Then

$$R(S_i)\zeta \geq R(S_j)\left(\mathbf{w}(j)-\sqrt{\mathbf{w}(j)^2-(2\mathbf{w}(i)\zeta+\zeta^2)}\right). \tag{30}$$

On the other hand

$$\begin{aligned}-\Delta f^2 &= R(S_i)(-\zeta)+R(S_j)\\ &\quad\times\left(-\mathbf{w}(j)+\sqrt{\mathbf{w}(j)^2-(-2\mathbf{w}(i)\zeta+\zeta^2)}\right)\end{aligned}$$

i.e.,

$$R(S_i)\zeta = \Delta f^2 - R(S_j)\mathbf{w}(j)+R(S_j)\sqrt{\mathbf{w}(j)^2-(-2\mathbf{w}(i)\zeta+\zeta^2)}.$$

Hence, by (30)

$$\begin{aligned}\Delta f^2 - R(S_j)&\left(\mathbf{w}(j)-\sqrt{\mathbf{w}(j)^2-(-2\mathbf{w}(i)\zeta+\zeta^2)}\right)\\ &\geq R(S_j)\left(\mathbf{w}(j)-\sqrt{\mathbf{w}(j)^2-(2\mathbf{w}(i)\zeta+\zeta^2)}\right)\end{aligned}$$

yielding

$$\begin{aligned}\Delta f^2 &\geq R(S_j)\left(2\mathbf{w}(j)-\sqrt{\mathbf{w}(j)^2-(2\mathbf{w}(i)\zeta+\zeta^2)}\right.\\ &\qquad\left.-\sqrt{\mathbf{w}(j)^2-(-2\mathbf{w}(i)\zeta+\zeta^2)}\right)\\ &\geq R(S_j)\mathbf{w}(j)\left(\left[1-\sqrt{1-\frac{2\mathbf{w}(i)\zeta+\zeta^2}{\mathbf{w}(j)^2}}\right]\right.\\ &\qquad\left.+\left[1-\sqrt{1+\frac{2\mathbf{w}(i)\zeta-\zeta^2}{\mathbf{w}(j)^2}}\right]\right).\end{aligned} \tag{31}$$

Then, using Taylor development

$$\begin{cases}\sqrt{1-\epsilon}=1-\frac{\epsilon}{2}-\frac{\epsilon^2}{8}+o(\epsilon^2)\\ \sqrt{1+\epsilon'}=1+\frac{\epsilon'}{2}-\frac{\epsilon'^2}{8}+o(\epsilon'^2)\end{cases} \tag{32}$$

where $o(\epsilon^2)$ and $o(\epsilon'^2)$ denote terms tending to zero faster than $||\epsilon^2||$ and $||\epsilon'^2||$, respectively. Hence, for sufficiently small ϵ and ϵ', one gets

$$\begin{cases} 1-\sqrt{1-\epsilon} > 1-\left(1-\frac{\epsilon}{2}\right) \\ 1-\sqrt{1+\epsilon'} > 1-\left(1+\frac{\epsilon'}{2}\right) \end{cases}. \tag{33}$$

Then, by letting $\epsilon = \zeta(2\mathbf{w}(i)+\zeta/\mathbf{w}(j)^2)$ and $\epsilon' = \zeta 2\mathbf{w}(i) - \zeta/\mathbf{w}(j)^2)$, we have for ζ small enough

$$\begin{cases} 1-\sqrt{1-\dfrac{2\mathbf{w}(i)\zeta+\zeta^2}{\mathbf{w}(j)^2}} > 1-\left(1-\dfrac{2\mathbf{w}(i)\zeta+\zeta^2}{2\mathbf{w}(j)^2}\right) \\ 1-\sqrt{1+\dfrac{2\mathbf{w}(i)\zeta-\zeta^2}{\mathbf{w}(j)^2}} > 1-\left(1+\dfrac{2\mathbf{w}(i)\zeta-\zeta^2}{2\mathbf{w}(j)^2}\right) \end{cases}. \tag{34}$$

By (34) and using inequality (31), it comes that for sufficiently small $\zeta > 0$

$$\begin{aligned} \Delta f^2 &> R(S_j)\mathbf{w}(j)\left(\frac{2\mathbf{w}(i)\zeta+\zeta^2}{2\mathbf{w}(j)^2} - \frac{2\mathbf{w}(i)\zeta-\zeta^2}{2\mathbf{w}(j)^2}\right) \\ &= \frac{R(S_j)\zeta^2}{\mathbf{w}(j)} > 0. \end{aligned}$$

■

Appendix V
Empirical Rule for Choosing a Default Value of m

We want to find $m^\star$ such that for all $m \le m^\star$, the estimation error is small (say less than $\mathcal{E}$) with a high probability (say higher than $\mathcal{L}(m^\star)$)

$$\Pr\left[R(X) - \langle R^\star_m(X)\rangle \le \mathcal{E}\right] \ge \mathcal{L}(m^\star) \tag{35}$$

where $\mathcal{L}(m^\star)$ is a probability threshold. The main problem of this approach is that if $\mathcal{E}$ is a constant, we are not able to find an expression for $\mathbf{L}(m^\star)$ that is useful and *blind*, that is *distribution-free* in the sense that it does not depends on f_X. For instance, the probability in (35) can be written as $1 - F_{\langle R^\star_m(X)\rangle}(R(X)-\mathcal{E})$, which depends on f_X. Thus, the point is to include the density dependency into the error term $\mathcal{E}$ [39]. Let us approximate the range measure by using quantile differences, and define the error term as

$$\mathcal{E}(X) \triangleq R(X) - (Q_X(q) - Q_X(p)) \tag{36}$$

where $0 \le p < q \le 1$ and $Q_X(.)$ is the quantile function defined as the inverse of the cdf: $Q_X(F_X(x)) = x$. Note that $\mathcal{E}(X)$ is positive and tends to 0 for increasing q and decreasing p, whatever is the density of X, but at a various rate. For example, with $q = .95$ and $p = 1-q$, we have $\mathcal{E}(T) = 31.6\%$ and $\mathcal{E}(V) = 5\%$ (see Fig. 3).

Observe that defining $\Gamma = R(X) - \mathcal{E}$, any lower bound of $\Pr[R^\star_m(X) \ge \Gamma]$ can be used in the right-hand side of (35)

$$\begin{aligned} \Pr\left[\langle R^\star_m(X)\rangle \ge \Gamma\right] &= \Pr\left[\langle R^\star_m(X)\rangle \ge \Gamma | R^\star_m(X) \ge \Gamma\right] \\ &\quad \times \Pr\left[R^\star_m(X) \ge \Gamma\right] \\ &\quad + \Pr\left[\langle R^\star_m(X)\rangle \ge \Gamma |^\star_m(X) < \Gamma\right] \\ &\quad \times \Pr\left[R^\star_m(X) < \Gamma\right] \\ &\ge \Pr\left[R^\star_m(X) \ge \Gamma\right] \end{aligned} \tag{37}$$

where the inequality results from the fact that $\langle R^\star_m(X)\rangle \ge R^\star_m(X)$ with probability one.

On the other hand, using the confidence interval for quantiles [40], noting that $\Pr[R^\star_m(X) \ge R^\star_{m^\star}(X) | m \le m^\star] = 1$ and setting $p = 1-q$ in (36), $\Pr[R^\star_m(X) \ge Q_X(q) - Q_X(1-q)] \ge \mathcal{L}(q, m^\star, N)$ for all $m \le m^\star$ with

$$\begin{aligned} \mathcal{L}(q, m^\star, N) \triangleq &\sum_{i=m^\star}^{N} \binom{N}{i} q^{N-i}(1-q)^i \\ &- \sum_{i=N-m^\star+1}^{N} \binom{N}{i} q^i (1-q)^{N-i} \end{aligned} \tag{38}$$

and, consequently, using inequality (37) and $\mathcal{E}(X)$ given by (36)

$$\Pr\left[R(X) - \langle R^\star_m(X)\rangle \le \mathbf{c}E(X)\right] \ge \mathcal{L}_+(q, m^\star, N) \tag{39}$$

with $\mathcal{L}_+(q, m^\star, N) \triangleq \max(0, \mathcal{L}(q, m^\star, N))$. The last inequality can be understood as follows: if q is chosen close enough to one, $\langle R^\star_m(X)\rangle$ *nearly covers* the true range, with a probability higher than $\mathcal{L}_+(q, m^\star, N)$. Note that q has to be chosen close enough to one, so that $\mathcal{E}(X)$ is small; otherwise, the bound $\mathcal{L}_+$ in (39) is no more related to range estimation quality. The terms *close enough to one* depends on the cdf F_X. In practice, however, if no information on the source densities is available, q can be *a priori* fixed to, e.g., 0.95 and $p = 1-q$. We take $m^\star$ as the largest value of m ensuring that $\mathcal{L}_+(q, m^\star, N)$ is greater than a fixed threshold close but smaller than one (typically, we search for $m^\star$ such that $\mathcal{L}_+(q, m, N) \ge 0.95$), for fixed q and $p = 1-q$. This choice guarantees that the left-hand side probability in (39) is also greater than the aforementioned threshold for all $m \le m^\star$. The single parameter m has thus been replaced by two parameters, but the proposed approach has two advantages. First, the new parameters have a concrete interpretation; q is related to the range estimation and the bound $\mathcal{L}_+$ tells us the confidence that we can have in the range estimation. Second, in practice, q and $\mathcal{L}_+$ can be fixed, so that a direct relation between $m^\star$ and N is found, which can be used to set a default value for $m^\star$.

In Fig. 7, we plot the maximum value of m, i.e., $m^\star$, so that the quantity $\mathcal{L}_+(q, m^\star, N)$ equals various fixed values (indicated on the related curve) with respect to N. Null values for $m^\star$ indicate that it does not exist $m^\star \in \mathbb{Z}^+$ such that $\mathcal{L}_+(.95, m^\star, N)$ is greater or equal to the associated threshold for fixed N. In other words, by transitivity of the inequality, each couple (m, N) located under these curves ensures that the left-hand side probability in (39) is greater than the previous threshold. Observe that for sufficiently large N, small m, and for a given q, $\mathcal{L}_+(q, m, N)$ tends to one.

To avoid numerical problems, we suggest the use of logarithms when computing the binomial coefficients, i.e., $\binom{N}{i} = \exp[\sum_{j=1}^{N}\log j - \sum_{j=1}^{N-i}\log j - \sum_{j=1}^{i}\log j]$. If one desires to speed up the method, the following empirical law is proposed for selecting a default value for m; we can take

$$m^\sharp(N) = \max\left(1, \left\lceil \Re\left\{\left(\frac{N-18}{6.5}\right)^{0.65}\right\} - 4.5 \right\rceil\right) \tag{40}$$

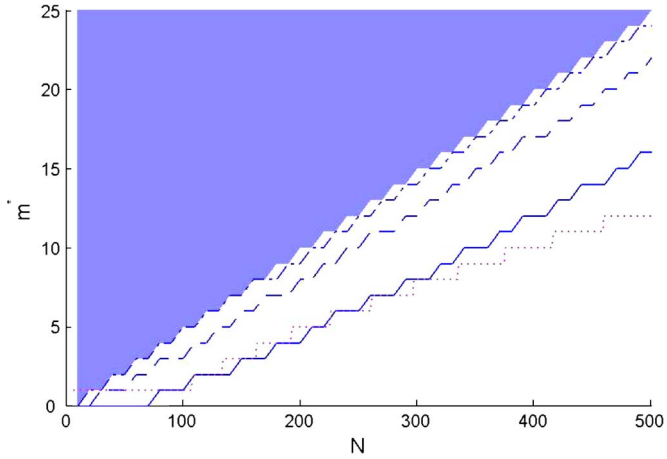

Fig. 7. Selected iso-$\mathcal{L}_+(q, m^\star, N)$ curves for $q = 0.95$ in the m versus N plane; $\mathcal{L}_+ = 0.1$ (dashed–dotted), $\mathcal{L}_+ = 0.5$ (dashed), and $\mathcal{L}_+ = 0.95$ (solid). The curve $m^\sharp$ given by (40) versus N has been also plotted (dotted). The "triangular" dark area indicates the set of points (m, N) for which $\mathcal{L}_+(0.95, m, N) = 0$ (useless bound: m must be out of this zone to ensure $\mathcal{L}_+(0.95, m, N) > 0$).

where $\overline{\alpha}$ denotes the nearest integer to α. This choice correspond to the dotted "step-like" curve in Fig. 7. Note that since $\mathcal{L}_+(q, m, N)$ decreases with m for fixed N and increases with N for fixed m, $\mathcal{L}_+(q, m^\sharp, N) \geq 0.95$ for $N > 210$.

ACKNOWLEDGMENT

The authors would like to thank C. Jutten from Institut National Polytechnique de Grenoble, France and D.-T. Pham from Institut d'Informatique et de Mathématiques Appliquées de Grenoble and Centre National de la Recherche Scientifique, France, for fruitful discussions and useful comments on a previous version of this paper. They would also like to thank the anonymous reviewers for their valuable remarks that led to the improvement of the clarity and the quality of this paper.

REFERENCES

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

[2] S. Haykin, Ed., *Unsupervised Adaptive Filtering, Vol. 1: Blind Source Separation*. New York: Wiley, 2000.

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

[4] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non-Gaussian signals," *Inst. Elect. Eng. Proc. F: Radar Signal Process.*, vol. 140, no. 6, pp. 362–370, 1993.

[5] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm independent component analysis," *Neural Comput.*, vol. 9, no. 7, pp. 1483–1492, 1997.

[6] Z. Koldovsky, D. Tichavsky, and E. Oja, "Efficient variant of algorithm fastICA for independent component analysis," *IEEE Trans. Neural Netw.*, vol. 17, no. 5, pp. 1265–1277, Sep. 2006.

[7] A. J. Bell and T. J. Sejnowski, "An information-maximisation approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, 1995.

[8] T.-W. Lee, M. Girolami, and T. J. Sejnowski, "Independent component analysis using an extended infomax algorithm for mixed sub-Gaussian and super-Gaussian sources," *Neural Comput.*, vol. 11, pp. 417–441, 1999.

[9] E. G. Learned-Miller and J. W. Fisher, III, "ICA using spacings estimates of entropy," *J. Mach. Learn. Res.*, vol. 4, pp. 1271–1295, 2003.

[10] L. Almeida, "Misep-linear and nonlinear ICA based on mutual information," *J. Mach. Learn. Res.*, vol. 4, pp. 1297–1318, 2003.

[11] R. Boscolo, H. Pan, and V. Roychowdhury, "Independent component analysis based on nonparametric density estimation," *IEEE Trans. Neural Netw.*, vol. 15, no. 1, pp. 55–65, Jan. 2004.

[12] A. Cichocki and S.-I. Amari, *Adaptive Blind Signal and Image Processing*. London, U.K.: Wiley, 2002.

[13] F. Theis, P. Georgiev, and A. Cichocki, "Robust overcomplete matrix recovery for sparse sources using a genralized Hough transform," in *Proc. Euro. Symp. Artif. Neural Netw. (ESANN)*, Bruges, Belgium, 2004, pp. 343–348.

[14] M. Plumbley, "Algorithms for nonnegative independent component analysis," *IEEE Trans. Neural Netw.*, vol. 4, no. 3, pp. 534–543, May 2003.

[15] L. Wei and J. Rajapakse, "Approach and applications of constrained ICA," *IEEE Trans. Neural Netw.*, vol. 16, no. 1, pp. 203–212, Jan. 2005.

[16] F. Vrins and M. Verleysen, "On the entropy minimization of a linear mixture of variables for source separation," *Signal Process.*, vol. 85, no. 5, pp. 1029–1044, 2005.

[17] D.-T. Pham and F. Vrins, "Local minima of information-theoretic criteria in blind source separation," *IEEE Signal Process. Lett.*, vol. 12, no. 11, pp. 788–791, Nov. 2005.

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

[19] J. Murillo-Fuentes and F. Gonzalez-Serrano, "A sinusoidal contrast function for the blmind separation of statistically independent sources," *IEEE Trans. Signal Process.*, vol. 52, no. 12, pp. 3459–3463, Dec. 2004.

[20] F. Theis, A. Jung, C. Puntonet, and E. Lang, "Linear geometric ICA: fundamentals and algorithms," *Neural Comput.*, vol. 15, pp. 419–439, 2003.

[21] A. Prieto, C. Puntonet, and B. Prieto, "A neural learning algorithm for blind separation of sources based on geometric properties," *Signal Process.*, vol. 64, pp. 315–331, 1998.

[22] A. Erdogan, "A simple geometric blind source separation method for bounded magnitude sources," *IEEE Trans. Signal Process.*, vol. 54, no. 2, pp. 438–447, Feb. 2006.

[23] D.-T. Pham, "Blind separation of instantaneous mixtures of sources based on order statistics," *IEEE Trans. Signal Process.*, vol. 48, no. 2, pp. 363–375, Feb. 2000.

[24] S. Cruces and I. Duran, "The minimum support criterion for blind source extraction: a limiting case of the strengthened Young's inequality," in *Lecture Notes in Computer Science*, ser. LNCS 3195, C. Puntonet and A. Prieto, Eds. Berlin, Germany: Springer-Verlag, Sep. 2004, pp. 57–64.

[25] F. Vrins, C. Jutten, and M. Verleysen, "SWM: a class of convex contrasts for source separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Philadelphia, PA, Mar. 2005, pp. V.161–V.164.

[26] J. Even, "Contributions à la séparation de sources à l'aide de statistiques d'ordre," Ph.D. dissertation, Univ. J. Fourier, Grenoble, France, 2003.

[27] Y. Blanco and S. Zazo, "An overview of BSS techniques based on order statistics: formulation and implementation issues," in *Lecture Notes in Computer Science*, ser. (LNCS 3195), C. Puntonet and A. Prieto, Eds. Berlin, Germany: Springer-Verlag, Sep. 2004, pp. 73–80.

[28] S. Fiori, "A theory for learning by weight flow on Stiefel-Grassman manifold," *Neural Comput.*, vol. 13, pp. 1625–1647, 2001.

[29] K. Hild, D. Erdogmus, and J. Principe, "Blind source separation using Renyi's mutual information," *IEEE Signal Process. Lett.*, vol. 8, no. 6, pp. 174–176, Jun. 2001.

[30] J. Principe, D. Xu, and J. Fisher, III, "Information-theoretic learning," in *Unsupervised Adaptive Filtering*. New York: Wiley, 2000, ch. I, pp. 265–319.

[31] O. Guleryuz, E. Lutwak, D. Yang, and G. Zhang, "Information-theoretic inequalities for contoured probability distributions," *IEEE Trans. Inf. Theory*, vol. 48, no. 8, pp. 2377–2383, Aug. 2002.

[32] F. Vrins, C. Jutten, D. Erdogmus, and M. Verleysen, "Zero-entropy minimization for blind extraction of bounded sources (BEBS)," in *Lecture Notes in Computer Science*, ser. LNCS 3889, J. Rosca, D. Erdogmus, J. Principe, and S. Haykin, Eds. Berlin, Germany: Springer-Verlag, Mar. 2006, pp. 747–754.

[33] M. Plumbley, "Lie group methods for optimization with orthogonality constraints," in *Lecture Notes in Computer Science*, ser. LNCS 3195, C. Puntonet and A. Prieto, Eds. Berlin, Germany: Springer-Verlag, 2004, pp. 1245–1252.

[34] C. Chefd'Hotel, D. Tschumperlé, R. Deriche, and O. Faugeras, "Regularizing flows for constrained matrix-valued images," *J. Math. Imag. Vis.*, vol. 20, pp. 147–162, Jan. 2004.

[35] L. Devroye and G. Wise, "Detection of abnormal behavior via nonparametric estimation of the support," *SIAM J. Appl. Math.*, vol. 38, pp. 480–488, 1980.
[36] J. Lee, F. Vrins, and M. Verleysen, "A simple ICA algorithm for nondifferentiable contrasts," in *Proc. Euro. Signal Process. Conf. (EUSIPCO)*, Antalya, Turkey, 2005, pp. cr1412.1–cr1412.4.
[37] F. Vrins, J. Lee, and M. Verleysen, "Filtering-free blind separation of correlated images," in *Lecture Notes in Computer Science*, ser. LNCS 3512, J. Cabestany, A. Prieto, and F. Sandoval, Eds. Berlin, Germany: Springer-Verlag, 2005, pp. 1091–1099.
[38] J. Lee, F. Vrins, and M. Verleysen, "A least absolute bound approach to ICA—application to the MLSP 2006 competition," in *Proc. IEEE Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2006, pp. 41–46.
[39] F. Vrins and M. Verleysen, "Minimum support ICA using order statistics. part I: quasi-range based support estimation," in *Lecture Notes in Computer Science*, ser. LNCS 3889, J. Rosca, D. Erdogmus, J. Principe, and S. Haykin, Eds. Berlin, Germany: Springer-Verlag, Mar. 2006, pp. 262–269.
[40] J. Chu, "Some uses of quasi-ranges," *Ann. Math. Statist.*, no. 28, pp. 173–180, 1957.
[41] F. Vrins, D.-T. Pham, and M. Verleysen, "Mixing and nonmixing local minima of the entropy contrast for blind source separtation," *IEEE Trans. Inf. Theory*, 2007, to be published.

Frédéric Vrins (S'06) was born in Uccle, Belgium, in 1979. He received the M.S. degree in mechatronics engineering and the D.E.A. degree in applied sciences from the Université catholique de Louvain (UCL), Louvain-la-Neuve, Belgium, in 2002 and 2004, respectively. He is currently working toward the Ph.D. degree in the UCL Machine Learning Group.

He is author or coauthor of more than 20 papers in international journals or conference proceedings with reviewing committee, and member of the program committee of ICA 2006. His research interests are blind source separation, independent component analysis, Shannon and Renyi entropies, mutual information, and information theory in adaptive signal processing.

John A. Lee was born in Brussels, Belgium, in 1976. He received the M.Sc. degree in applied sciences (computer engineering) in 1999 and the Ph.D. degree in applied sciences (machine learning) in 2003 from the Université catholique de Louvain (UCL), Louvain-la-Neuve, Belgium.

His main interests are nonlinear dimensionality reduction, intrinsic dimensionality estimation, independent component analysis, clustering, and vector quantization. He is a former member of the UCL Machine Learning Group and is now a Postdoctoral Researcher of the Belgian Fonds National de la Recherche Scientique (F.N.R.S.). His current work aims at developing specific image enhancement techniques for positron emission tomography in the Molecular Imaging and Experimental Radiotherapy Department, Saint-Luc University Hospital, Brussels, Belgium.

Michel Verleysen (S'87–M'92–SM'04) was born in Belgium in 1965. He received the M.S. and Ph.D. degrees in electrical engineering from the Université catholique de Louvain, Louvain-la-Neuve, Belgium, in 1987 and 1992, respectively.

He was an Invited Professor at the Swiss Ecole Polytechnique Fédérale de Lausanne (E.P.F.L.), Switzerland, in 1992, at the Université d'Evry, Val d'Essonne, France, in 2001, and at the Université Paris IPanthéon-Sorbonne, Paris, France, in 2002, 2003, and 2004, respectively. He is now a Professor at the Université catholique de Louvain and the Honorary Research Director of the Belgian National Fund for Scientific Research (F.N.R.S.) and a Lecturer at the Université catholique de Louvain. He is author or coauthor of about 200 scientific papers in international journals and books or communications to conferences with reviewing committee. He is the coauthor of the scientific popularization book on artificial neural networks in the series "Que Sais-Je?", in French. His research interests are in artificial neural networks, self-organization, time-series forecasting, nonlinear statistics, adaptive signal processing, information-theoretic learning, and biomedical data and signal analysis.

Dr. Verleysen is the Editor-in-Chief of the *Neural Processing Letters*, the Chairman of the annual European Symposium on Artificial Neural Networks Conference (ESANN), the Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS, and member of the editorial board and program committee of several journals and conferences on neural networks and learning.