Neurocomputing 72 (2008) 47–56

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Blind source separation based on endpoint estimation with application to the MLSP 2006 data competition

John A. Lee [a,b,*,1], Frédéric Vrins [c,d], Michel Verleysen [c]

[a] Belgian Fund for Scientific Research (FNRS), Machine Learning Group, Université catholique de Louvain, Place du Levant 3, B-1348 Louvain-la-Neuve, Belgium
[b] Center for Molecular Imaging and Experimental Radiotherapy, Université catholique de Louvain, Avenue Hippocrate 54, B-1200 Bruxelles, Belgium
[c] Microelectronics Laboratory, Machine Learning Group, Université catholique de Louvain, Place du Levant 3, B-1348 Louvain-la-Neuve, Belgium
[d] ING South West Europe, Financial Markets Department (Credit Trading Desk) Avenue Marnix 24, Marnix II+4, 1000 Brussels, Belgium

ARTICLE INFO

Available online 27 September 2008

Keywords:
Blind source separation
Independent component analysis
Range estimation
Endpoint estimation
Contrast function
Discriminacy
Deflation

ABSTRACT

The problem of blind source separation is usually solved by optimizing a contrast function that measures either the independence of several variables or the non-gaussianity of a single variable. If the problem involves bounded sources, this knowledge can be exploited and the solution can be found with a customized contrast that relies on a simple endpoint estimator. The minimization of the least absolute endpoint is closely related to and generalizes the minimization of the range, which has already been studied within the framework of blind source extraction. Using the least absolute endpoint rather than the range applies to a broader class of admissible sources, which includes sources that are bounded on a single side and, therefore, have an infinite range. This paper describes some properties of a contrast function based on endpoint estimation, such as the discriminacy. This property guarantees that each local minimum of the least absolute bound corresponds to the extraction of one source. An endpoint estimator is proposed, along with a specific deflation algorithm that is able to optimize it. This algorithm relies on a loose orthogonality constraint that reduces the accumulation of errors during the deflation process. This allows the algorithm to solve large-scale and ill-conditioned problems, such as those proposed in the MLSP 2006 data competition. Results show that the proposed algorithm outperforms more generic source separation algorithms like FastICA, as the sources involved in the contest are always bounded on at least one side.

1. Introduction

Blind source separation (BSS) has proved to be useful in many areas of signal processing. Applications are numerous in denoising [7], acoustics [9], biomedical signal analysis [13], and in other domains that involve arrays of sensors. However, in most of these applications, general-purpose methods are used. These methods, such as JADE [2] and FastICA [12,10], are widely available and have been well studied in the literature (see [11] and references therein). They also provide the user with an interesting tradeoff between ease of use and overall good performance.

In practice, the BSS problem can be solved only if the sources and the mixtures fulfill some conditions. The most common assumptions are that the sources are statistically independent, and that the mixtures are linear and instantaneous. This framework correspond to independent component analysis [11] (ICA). Finding a solution to the ICA problem requires the definition of a contrast function, namely an objective function that is applied to one or several linear combinations of the mixtures, and whose global maxima correspond to desired solutions of the problem.

This paper aims at studying a contrast function that relies on some *a priori* assumptions about the sources signals. If these assumptions are valid, an algorithm based on this specific contrast function can dramatically increase the separation performance, compared to usual techniques. The literature gathers several works that develop similar approaches. For instance, assumptions such as the sparsity [31] or the non-negativity [20] of the sources have been investigated. Our approach is related to previous works about the range of the estimated sources [28,29,25], also known as the support width measure (SWM) [23,15]. As a contrast function, the range has interesting properties, such as the absence of local minima [26]; it can even be used with dependent sources, provided all sources have a bounded support [25].

We extend this approach to sources whose supports are bounded only on one side, such as non-negative sources, among other possibilities. For this purpose, we assume that sources are

[*] Corresponding author at: Center for Molecular Imaging and Experimental Radiotherapy, Université catholique de Louvain, Avenue Hippocrate 54, B-1200 Bruxelles, Belgium. Tel.: +32 2 7644778; fax: +32 10 47 25 98.
E-mail addresses: John.Lee@uclouvain.be (J.A. Lee), Frederic.Vrins@ing.be (F. Vrins), Michel.Verleysen@uclouvain.be (M. Verleysen).

[1] Postdoctoral Researcher of the Belgian National Fund of Scientific Research (FNRS).

0925-2312/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2007.12.047

centered and we replace the range with an estimator of the lowest endpoint, in absolute value. An important result is that we show that the proposed contrast keeps the discriminacy property, i.e. to each of its local maxima corresponds a linear combination of the mixtures that actually returns a single source, up to some scaling and permutation.

As the least absolute endpoint (LAE) is not everywhere differentiable, we propose a specific algorithm for its minimization. Each part of the algorithm (whitening, deflation, loose orthogonality constraint) have been designed in order to maximize the separation performance in difficult problems. Experiments detailed in this paper come from the "ICA algorithms for large-scale, ill-conditioned, and noisy mixtures" data analysis competition at the IEEE MLSP 2006 workshop. The proposed technique has won this contest.

The remainder of this paper is organized as follows: Section 2 introduces ICA. Section 3 describes the proposed contrast, which is based on the LAE. Section 4 describes an estimator of the contrast, along with a specific optimization procedure. Section 5 details the results obtained with the mixtures proposed in the IEEE MLSP 2006 competition. Finally, conclusions are drawn in Section 6.

2. The framework of ICA

The problem of BSS is usually tackled within the framework of ICA. The main hypotheses in ICA are the independence and stationarity of the sources. As for the mixtures, one usually assumes that they are linear and instantaneous. If $\mathbf{s} = [s_1, \ldots, s_m]^T$ denotes the vector of unknown sources, then the mixture model can be written as $\mathbf{x} = \mathbf{A}\mathbf{s}$, where $\mathbf{x} = [x_1, \ldots, x_n]^T$ is the vector of observed mixtures and $\mathbf{A}$ is the unknown mixing matrix. Throughout this paper, we assume that the number of mixtures is equal to the number of sources ($m = n$).

2.1. Contrast function

Statistical independence is the key assumption that allows ICA to recover the unknown sources. In practice, almost all ICA algorithms actually achieve the optimization of a quantity that measures in some way the independence of the estimated sources. Informally, a function is thus a contrast if its maximization allows us to identify either one or all sources. This leads to the distinction of two classes of contrasts.

In the first class, the contrast is a function $\mathscr{C}$ of all estimated sources. The essential property of such a contrast is that it reaches a global maximum if and only if the estimated sources correspond to the actual ones, up to a scaling and a permutation. If the estimated sources are denoted $\mathbf{y} \doteq \mathbf{B}\mathbf{x}$, then $\arg\max_{\mathbf{B}} \mathscr{C}(\mathbf{y})$ identifies the demixing matrix (up to the above indeterminancies) and leads to the simultaneous separation of all sources. Algorithms that perform this maximization are often said to be symmetric. Estimators of the mutual information [30,19] belong to this first class.

In the second class, the contrast is applied to a single estimated source and it reaches a maximum when the estimated source corresponds to one of the actual sources. If $\mathbf{y} = [y_1, \ldots, y_m]^T$, then $y_i = \mathbf{b}\mathbf{x}$ and $\arg\max_{\mathbf{b}} \mathscr{C}(y_i)$ identifies a row of the demixing matrix $\mathbf{B}$. Notice that k distinct maxima have to be found in order to extract k sources from the set of mixtures. For this reason, algorithms that extract sources one by one by repeating the contrast maximization are said to follow a deflation procedure. The kurtosis [16] is a typical example of a deflation contrast.

Of course, the fact that a contrast has to be maximized is purely conventional. In the cases of the mutual information and the range, it is often simpler to minimizes those quantities instead of introducing minus signs all around. Notice that, although the definition of a contrast guarantees that the solution of the BSS problem corresponds to maxima of the contrast, the converse statement does not necessarily hold true. This means that, depending on the contrast, local maxima can exist, which do not correspond to any solution of the problem. Contrasts that are free of spurious maxima are said to be discriminant [25].

2.2. Whitening

Many ICA algorithms comprise a whitening step that simplifies the separation problem. Prewhitened mixtures are denoted by $\mathbf{z} \doteq \mathbf{V}\mathbf{x}$, where the whitening matrix $\mathbf{V}$ is such that $\mathrm{E}\mathbf{z}$ and the covariance matrix $\mathrm{E}\{\mathbf{z}\mathbf{z}^T\}$ is equal to the identity and the product $\mathbf{V}\mathbf{A}$ is orthogonal. After this preprocessing, ICA reduces to finding an orthogonal matrix $\mathbf{W}$ and $\mathbf{y} = \mathbf{W}\mathbf{z}$ yields an estimate of the sources up to a scaling and permutation. As ICA cannot recover the variance of the original sources [11], this means that each output will correspond to a whitened source (zero-mean, unit-variance), whose sign is not relevant. In other words, the output vector $\mathbf{y} = \mathbf{W}\mathbf{V}\mathbf{A}(\mathbf{s} - \mathrm{E}\{\mathbf{s}\})$ is such that $\mathrm{E}\{\mathbf{y}\mathbf{y}^T\} = \mathbf{I}$. In order to compensate for centering, the vector $\mathbf{B}\mathrm{E}\{\mathbf{x}\} = \mathbf{W}\mathrm{E}\{\mathbf{z}\}$ can always be added to $\mathbf{y}$. This shows that without loss of generality, the sources can be assumed to be white in the identification of $\mathbf{B}$, for the sake of simplicity. Under this hypothesis, the product $\mathbf{V}\mathbf{A}$ is orthogonal and such is also the transfer matrix $\mathbf{C} = \mathbf{B}\mathbf{A}$. Furthermore, we can assume that $\mathbf{B}$ and $\mathbf{A}$ are orthogonal matrices. In particular, in the two-dimensional case, the mixing matrix $\mathbf{A}$, the demixing matrix $\mathbf{B}$, and the transfer matrix $\mathbf{C}$ are rotation matrices with angles ϕ, φ, and $\theta = \varphi + \phi$, respectively. In this case, source separation amounts to finding φ such that θ is an entire multiple of $\pi/2$.

The above discussion shows that prewhitening simplifies the ICA problem from a purely theoretical point view, by reducing the space of possible solutions. In practice, however, whitening can be difficult to achieve in some cases, depending on the actual value of the mixing matrix $\mathbf{A}$. If $\mathbf{A}$ is large and has a high condition number, whitening can fail to decorrelate all mixtures. For this reason, the implementation of the prewhitening step must be carefully designed. As whitening is basically intended to decorrelate the mixtures, its accuracy also critically depends on the sample covariance matrix. If the sample size is low, or if there are outliers, whitening can thus be inaccurate.

Deflation algorithms that rely on a prewhitening step are often said to provide poor separation performances. For a high number of mixtures, this is mainly due to the accumulation of inaccuracies during the deflation process. As rows of $\mathbf{W}$ are related to each other with an orthogonality constraint, inaccuracies of a given row easily propagate to the subsequent ones. This issue can be addressed by relaxing the orthogonality constraint.

3. A contrast function based on endpoint estimation

Many general-purpose ICA algorithms, like FastICA [12,10] or JADE [2] offer an appealing tradeoff between separation performances and speed. In some demanding applications, however, any *a priori* knowledge about the sources can be taken into account in order to design a specific contrast that improves the performances. For instance, contrast functions for sparse [31] or non-negative [20] sources can be found in the literature. More generally, the improvement can impact at least four points:

- The speed and/or the convergence rate.
- The separation quality, measured with performance index such as the signal-to-interference ratio (SIR).

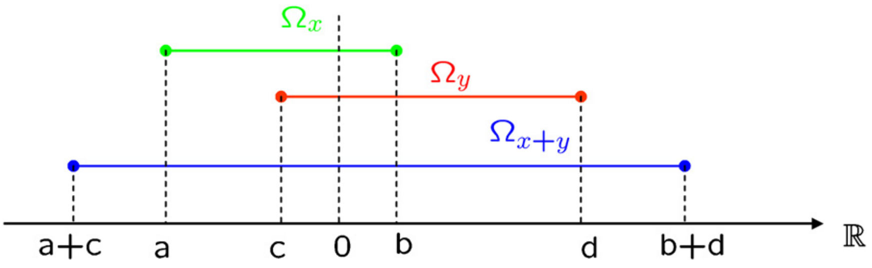

Fig. 1. Example of supports of zero-mean independent random variables x, y and the support of the sum of these variables.

- The possibility of relaxing some assumption in the model (such as the statistical independence of the sources or the usual requirement $m \leqslant n$).
- The possibility of designing a contrast function with more interesting properties, such as the discriminacy [25].

If all sources have upper and lower bounds, an algorithm that minimizes the SWM can address the last three points, depending on the context [28,29,24,15,23]. The algorithm described in [14] can extract sources one by one (using a deflation approach [3,23]) by minimizing the function $\mathscr{C}_\Omega(y_i) \doteq \sup \Omega_{y_i} - \inf \Omega_{y_i}$, where Ω_x denotes the support of the random variable x with the probability density function p_x: $\Omega_x \doteq \{\xi : p_x(\xi) > 0\}$. Thus, the function $\mathscr{C}_\Omega(y_i)$ is nothing but the range of the estimated source y_i.

In order to deal with a larger set of possible source distributions, we propose to replace the range estimation with an endpoint estimation [8]. This approach can be related to Erdogan's work [5], which deals with the minimization of the supremum for symmetric bounded signals. In our approach, symmetry is no longer required.

In statistics, $\theta(x)$ usually denotes the endpoint [6] of a random variable x. More precisely, if x has probability density function p_x and support $\Omega_x =]-\infty, a]\bigcup_{i=1}^{q}\{x_i\}$ for some finite scalar number a, and possible isolated points $x_1, \ldots, x_q$, with $q < \infty$, then $\theta(x) = \sup \Omega_x$. Assuming that x has a lower bound instead of an upper one, an equivalent definition involving the infimum of the support can be derived.

Compared to the range, the endpoint allows us to extract sources that are bounded only on a single side. Obviously, double-bounded sources can be extracted as well, provided we focus on one of their two endpoints. If the sources are centered, we can select the smallest endpoint in absolute value, for instance. The same trick is also useful for single-bounded sources, as ICA can only recover them up to a sign change. Consequently, we can define the LAE $\vartheta(x)$ of a random variable x with arbitrary support Ω_x as

$$\vartheta(x) \doteq \min\{|\inf \Omega_x|, |\sup \Omega_x|\}. \tag{1}$$

In the case of a single-bounded source, this guarantees that $\vartheta(\alpha x) < \infty$. The LAE also benefits from the properties stated in the following lemma.

Lemma 1 (*Properties of the LAE*). *Let x and y be two independent random variables. Then, the following properties hold:*

$\mathscr{P}_1$ *non-negativity*: $\vartheta(x) \geqslant 0$,
$\mathscr{P}_2$ *scale-equivariance*: $\vartheta(\alpha x) = |\alpha|\vartheta(x)$ *for all* $\alpha \in \mathbb{R}$,
$\mathscr{P}_3$ *superadditivity*: *if* $\mathrm{E}\{x\} = \mathrm{E}\{y\} = 0$, *then* $\vartheta(x+y) \geqslant \vartheta(x) + \vartheta(y)$.

The proof of this lemma is trivial and is not detailed here. The last property $\mathscr{P}_3$ is illustrated in Fig. 1 in the two-dimensional case. Assuming that $b \leqslant |a|$ and $|c| \leqslant d$, we have $\vartheta(x) = b, \vartheta(y) = |c|$, $b + |c| \leqslant \min\{|a| + |c|, b + d\}$, i.e. $\vartheta(x) + \vartheta(y) \leqslant \vartheta(x+y)$; equality is reached only if either $d = -c$ or $a = -b$, that is, if the support of at least one of the two sources is symmetric with respect to zero.

It is important to stress that $\vartheta(x+y)$ could take infinite values even if $\max\{\vartheta(x), \vartheta(y)\} < \infty$. This occurs for example if $\Omega_x = [a, \infty[$ and $\Omega_y = [-\infty, b[$ for some $-\infty < a, b < \infty$. In the following, we only assume that the sources have a finite LAE.

Based on Lemma 1, we obtain the following corollary, provided the sources are centered (this condition is trivially fulfilled if the mixtures are prewhitened).

Corollary 1 (*LAE contrast*). *With the above notations, the criterion*

$$\mathscr{C}_{\mathrm{LAE}}(y_i) \doteq \vartheta\left(\sum_{j=1}^{m} w_{ij} z_j\right) \tag{2}$$

is a contrast function for source separation by deflation.

Proof. Let us define by $\mathscr{I}(\mathbf{w})$ the set of indices j in $l, \cdots, m$ corresponding to the non-zero entries of the vector $\mathbf{w}$ in $\mathbb{R}$: $\mathscr{I}(\mathbf{w}) \doteq \{j : \mathbf{w}(j) \neq 0\}$. Noting further by $\mathbf{c}_i$ the ith row of $\mathbf{C}$, c_{ij} its entries, and $k \doteq \arg\min_i \vartheta(s_i)$, Lemma 1 allows us to write

$$\begin{aligned}
\vartheta^2(y_i) &= \vartheta^2(\mathbf{c}_i\mathbf{s}) \\
&\overset{\mathscr{P}_2,\mathscr{P}_3}{\geqslant} \sum_{j \in \mathscr{I}(\mathbf{c}_i)} c_{ij}^2 \vartheta^2(s_j) \\
&= \vartheta^2(s_k) + \sum_{j \in \mathscr{I}(\mathbf{c}_i)\setminus\{k\}} c_{ij}^2(\vartheta^2(s_j) - \vartheta^2(s_k)) \\
&\geqslant \vartheta^2(s_k),
\end{aligned}$$

where the equality results from the fact that $\|\mathbf{c}_i\| = 1$ for all $1 \leqslant i \leqslant n$.

The above inequalities are strict unless there is a single element in the set $\mathscr{I}(\mathbf{c}_i)$. Consequently, because the quadratic function is monotonously increasing on $\mathbb{R}^+$, the minimization of $\mathscr{C}_{\mathrm{LAE}}(y_i) = \vartheta(\sum_{j=1}^m w_{ij} z_j)$ with respect to $\mathbf{w}_i$ on the unit-sphere leads to the recovery of the source with the LAE, namely s_k. If we assume that the sources are indexed according to their LAE (i.e. $\vartheta(s_1) \leqslant \vartheta(s_2) \leqslant \cdots \leqslant \vartheta(s_m)$), we can iterate this result. Then the output signal $y_j = \sum_{i=1}^m w_{ji} z_i$ obtained by minimizing $\mathscr{C}_{\mathrm{LAE}}(y_j)$ with respect to $\mathbf{w}_j$ and subject to $\mathbf{w}_i \mathbf{w}_j^{\mathrm{T}} = 1$ for all $1 \leqslant i \leqslant j$ is a whitened copy of s_j (or any other source with the same LAE value). □

More generally, by using Pham's theorem about contrast functions [18], we find the following result: if mixtures are prewhitened, then $\mathbf{W}$ is constrained to belong to the group of orthogonal matrices,[1] and therefore $\prod_{i=1}^m \vartheta(\mathbf{w}_i \mathbf{z})$ is a contrast function for separating the m sources simultaneously. More generally, if mixtures are not prewhitened, then $\prod_{i=1}^m \vartheta(\mathbf{b}_i \mathbf{x}) / |\det \mathbf{B}|$ is a contrast function.

Fig. 2 illustrates that function $\mathscr{C}_{\mathrm{LAE}}(\cdot)$ fulfills the conditions to be a deflation contrast. Notice that in the two-dimensional case, the second source is trivially recovered at the same time as the

[1] As we deal with zero-mean signals, the fact that $\vartheta(\cdot)$ is not shift-invariant does not matter (i.e. for $\beta \in \mathbb{R}_0$, the equality $\vartheta(x + \beta) = \vartheta(x)$ does not necessarily hold).

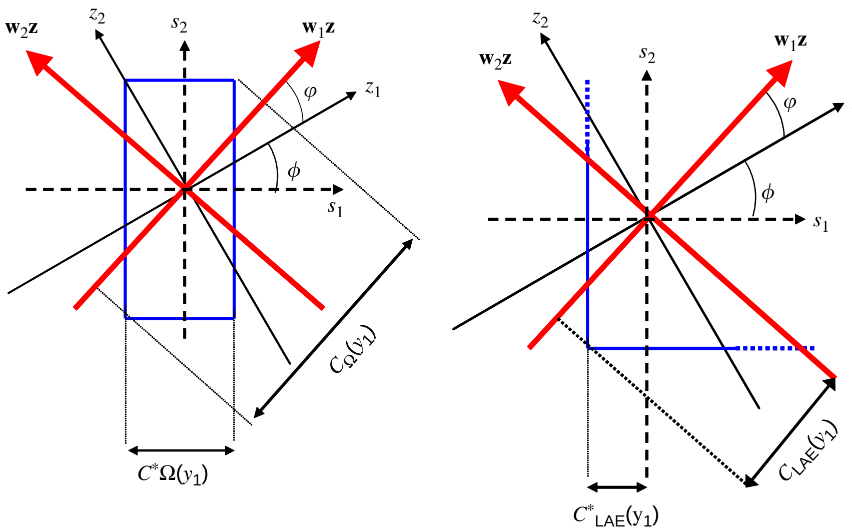

Fig. 2. Graphical interpretation of the SWM (left: two two-side bounded sources, $\mathcal{C}_\Omega$) and LAE (right: scatter two one-side bounded source, $\mathcal{C}_{\text{LAE}}$) in the plane (s_1, s_2); the minimum values of the contrasts are labelled "$\star$". The solid lines represents the support boundaries and the bold solid arrows represent the current axes.

first one, thanks to the orthogonality constraint that results from prewhitening. In this case, **A** and **B** are orthogonal, and source separation is achieved by finding a single angle.

3.1. Discriminacy property of the LAE contrast

A contrast function owns the property of discriminacy if it reaches a local minimum (or maximum, depending on the convention) only at a solution point [25]. In other words, finding any local minimum amounts to extracting one or several sources and, depending on the separation scheme (symmetric, by deflation, or partial), all are potentially relevant. This property is convenient as it turns a global optimization problem into a local one. Hence, the contrast function has no spurious optimum.

There are few discriminant contrast functions. To our knowledge, only kurtosis-based and range-based functions own this property [26]. As a counter example, contrast that are based on Shannon's entropy, including mutual-information and maximum negentropy method, might not be discriminant in specific cases [27].

Proving the discriminacy of the proposed contrast function is not easy because it contains operators such as min, inf, and sup, which are not everywhere differentiable. Additionally, the contrast is not necessarily strictly superadditive (the strict equality in $\mathscr{P}_3$ of Lemma 1 does not necessarily hold true). Therefore, we will first sketch some results about the contrast discriminacy in the two-dimensional case, with prewhitened mixtures and orthogonal demixing matrix **B**. To this end, we shall consider the example given in Fig. 1 (with $\Omega_{s_1} = \Omega_x$ and $\Omega_{s_2} = \Omega_y$). The extension to the multi-dimensional case will come next. The following lemma focuses on the contrast discriminacy in the case of a deflation procedure.

Lemma 2 (*Discriminacy of $\vartheta(\mathbf{wz})$ s.t. $\|\mathbf{w}\| = 1$ and $\mathbf{w} \in \mathbb{R}^2$*). *Assuming that $\mathbf{w} \in \mathbb{R}^2$ satisfies constraint $\|\mathbf{w}\| = 1$, any local minimum point of $\vartheta(\mathbf{wz})$ necessarily corresponds to* **wVA** *being proportional to a basis vector (i.e. proportional to either* $[0, 1]$ *or* $[1, 0]$*).*

Proof. Because **z** is a white random vector and **w** is a unit-norm vector, the output $y = \mathbf{wz}$ can be written as $y_\theta = s_1 \cos\theta + s_2 \sin\theta$. Actually, it is sufficient to prove that $\vartheta(y_\theta)$ is discriminant in a particular quadrant $\mathcal{Q}_i \doteq](i-1)\pi/2, i\pi/2[$, with $(i \in \mathbb{N}^*)$; this will extend to the other quadrants of the unit circle. Here, we consider only $\mathcal{Q}_1$. The discriminacy of $\vartheta(y_\theta)$ results from its "quadrant-wise" concavity. Indeed, it is the minimum of two functions $f_i(\theta)$ and $g_i(\theta)$ that are easily seen to be concave in $\mathcal{Q}_1$:

$$\vartheta(y_\theta) = \min\{\underbrace{|a|\cos\theta + |c|\sin\theta}_{\doteq f_1(\theta)}, \underbrace{b\cos\theta + d\sin\theta}_{\doteq g_1(\theta)}\}. \qquad (3)$$

Hence, $\vartheta(y_\theta)$ is also concave in $\mathcal{Q}_1$ (and actually in any quadrant of the unit circle, up to ad-hoc modifications in functions $f_i(\theta), g_i(\theta)$). By corollary, its local minima are necessarily attained at the borders of the quadrants, i.e. for $\theta \in \{k\pi/2, k \in \mathbb{Z}\}$.

Notice that depending on the values of a, b, c, d, the couple of curves $f_i(\theta)$ and $g_i(\theta)$ may or may not have an intersection in a given quadrant. For instance, it is obvious that the intersection of $f_1(\theta)$ and $g_1(\theta)$, if it exists, is unique and occur at the angle $\theta^\star(\mathcal{Q}_1) \doteq \arctan((|a| - b)/(d - |c|))$. For example, no such intersection exists in $\mathcal{Q}_1$ if $b > |a|$ and $d > |c|$ or $b < |a|$ and $d < |c|$. □

The above lemma is illustrated in Fig. 3 for four different quadruples (a, b, c, d) (see Fig. 1). The first and second quadrants $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are considered for completeness (the two others do not provide further information because of symmetry properties). Note that we have $f_2(\theta) = b|\cos\theta| + |c|\sin\theta$ and $g_2(\theta) = |a\cos\theta| + d\sin\theta$. Let us now extend Lemma 2 to $m > 2$.

Lemma 3 (*Discriminacy of $\vartheta(\mathbf{wz})$ s.t. $\|\mathbf{w}\| = 1$ and $\mathbf{w} \in \mathbb{R}^m$*). *After a prewhitening step, the contrast function $\vartheta(\mathbf{w}_i\mathbf{z})$ is discriminant on the set $\mathbf{w}_i \in \mathscr{S}^m$, where $\mathscr{S}^m$ is the m-dimensional unit-sphere*

$$\mathscr{S}^m \doteq \{\mathbf{w} \in \mathbb{R}^m : \|\mathbf{w}\| = 1\}.$$

Proof. The proof is based on the similarity between the two arguments of the min function involved in the definition of $\vartheta(x)$ (see Eq. (1)) and the range of x, defined as

$$R(x) \doteq \sup\Omega_x - \inf\Omega_x.$$

Thanks to prewhitening, one has $\mathrm{E}\{y_i\} = 0$ (and one can freely assume $\mathrm{E}\{s_j\} = 0$ for all j), $|\inf\Omega_{y_i}| = -\inf\Omega_{y_i}$, $|\sup\Omega_{y_i}| = \sup\Omega_{y_i}$, and $R(y_i)$ can all be written as a weighted sum of positive quantities depending on the sources and possibly the sign of the

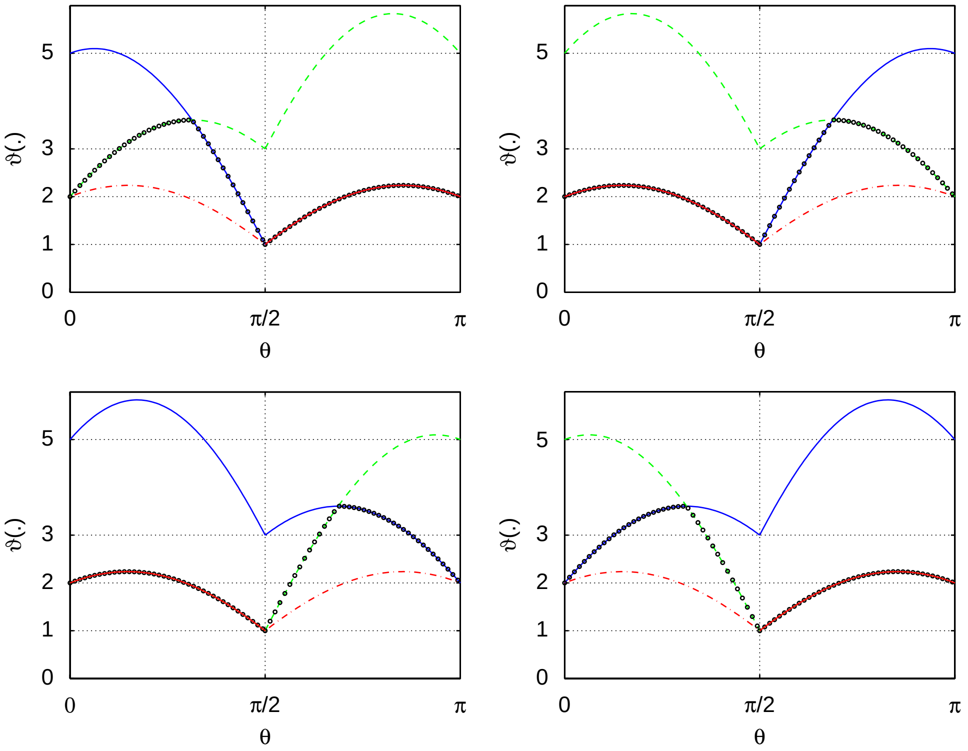

Fig. 3. Concavity of $\vartheta(y_\theta)$ where $y_\theta = \cos\theta s_1 + \sin\theta s_2$ with $a = \inf \Omega_{s_1}, b = \sup \Omega_{s_1}, c = \inf \Omega_{s_2}, d = \sup \Omega_{s_2}$. The functions $f_i(\theta)$ (solid) and $g_i(\theta)$ (dashed) are plotted in the two first quadrants for different quadruples (a, b, c, d) (clockwise from top left: $(-5, 2, -1, 3), (-2, 5, -1, 3), (-5, 2, -3, 1), (-2, 5, -3, 1)$). The $\vartheta(y_\theta)$ curve (with "∘" markers) is seen to be "quadrant-wise" concave (i.e. concave in each of the quadrants). Hence, all the local minimum points are located on the borders of the quadrants, meaning that the contrast is discriminant. Finally, one can check that $\vartheta(\cos\theta s_1) + \vartheta(\sin\theta s_2)$ (dash-dotted) is always lower than or equal to $\vartheta(y_\theta)$, as claimed in Lemma 1.

corresponding source weights:

$$-\inf \Omega_{y_i} = \sum_j |c_{ij}|(-\inf(\text{sign}(c_{ij})s_j)),$$

$$\sup \Omega_{y_i} = \sum_j |c_{ij}| \sup(\text{sign}(c_{ij})s_j),$$

$$R(y_i) = \sum_j |c_{ij}| R(s_j). \tag{4}$$

Observe that one must keep $\text{sign}(c_{ij})$ within the $\inf(\cdot)$ as a negative sign of c_{ij} would transform $\inf(\text{sign}(c_{ij})s_j)$ in $\sup(s_j)$, and conversely for the supremum.

Let us now focus on the part of the m-dimensional unit-sphere $\mathscr{S}^m$ belonging to the same hyper-quadrant as a given vector $\mathbf{c}_i$ satisfying $c_{ij}c_{ik} \neq 0$ for some indices j, k. This restriction, noted $R(\mathbf{c}_i)$, is the set of unit-norm m-dimensional vectors having the same sign vector as $\mathbf{c}_i$ (with $\text{sign}(0) = 0$ by convention). Mathematically, this restriction can be expressed as

$$\mathscr{R}(\mathbf{c}_i) \doteq \{\mathbf{w} \in \mathscr{S}^m\} \quad \text{s.t. } \forall j \ \text{sign}(w_j) = \text{sign}(c_{ij}),$$

which is defined for unitary vectors $\mathbf{w} \in \mathbb{R}^m \backslash \mathbb{I}^m$ only, $\mathbb{I}^m$ being the set of m-dimensional vectors having a single non-zero entry.

Assume $y_i = \mathbf{c}_i \mathbf{s}$ does not correspond to a source (i.e. $\mathbf{c}_i \notin \mathbb{I}^m$). It makes then sense to compute the restriction $\mathscr{R}(\mathbf{c}_i)$. On this restriction, each right-hand side of Eq. (4) is a weighted sum of constant, positive quantities (e.g. α_j^2) and can be written as

$$f(\alpha, \mathbf{c}_i) = \sum_j |c_{ij}| \alpha_j^2.$$

On the other hand, earlier works about the range-based contrast functions show that if $\alpha_j^2 > 0$ for all j, then f is free from local minimum point on the restriction $\mathscr{R}(\mathbf{c}_i)$ (see details in [22]).

From these results, one gets that neither $-\inf \Omega_{y_i}$ nor $\sup \Omega_{y_i}$ can have a local minimum point on $R(\mathbf{c}_i)$ (whatever is $\mathbf{w} \in \mathbb{R}^m \backslash \mathbb{I}^m$). Finally, it is easily seen that, consequently, $\vartheta(y_i)$, which is then the minimum of two functions free from local minimum on any restrictions $\mathscr{R}(\mathbf{c}_i)$, is itself free from local minimum on any restrictions $\mathscr{R}(\mathbf{c}_i)$, i.e. whatever is $\mathbf{c}_i \in \mathscr{S}^m$ provided that there exists j, k such that $c_{ij}c_{ik} \neq 0$.

Since this result does not depend on the number of nonzero entries of $\mathbf{c}_i$, this means that provided that there are at least two nonzero entries, one can set one additional entry to zero without facing a local minimum point. This concludes the proof of the discriminacy of $\vartheta(y_i)$. □

Notice that the symmetry of the (bounded) source densities is not required in the above lemma, even though it is a sufficient condition for the lemma to hold true.

4. Implementation

An ICA algorithm relies on several implementation choices. Mixtures can be prewhitened or processed as such. As for the contrast function, a good estimator that can be applied to finite samples must be found. Depending on the contrast nature, the algorithm can follow a deflation procedure or separate all sources jointly. The contrast optimization itself can work according different schemes, such as gradient descent or fixed-point iterations.

The algorithm we propose is intended to solve the problems of the MLSP 2006 competition. These problems involve a high number of mixtures, low sample sizes, and/or ill-conditioned mixtures of sources that are single-bounded or double-bounded. As the LAE is a property of scalar random variables, the algorithm naturally adopts a deflation architecture. We also developed a symmetric algorithm based on the LAE but only a few preliminary experiments sufficed to show that it scales poorly with the number of mixtures. This can easily be explained by the fact that symmetric algorithms have to update and orthonormalize the complete matrix $\mathbf{W}$ at each update, whereas deflation algorithms follow a "divide and conquer" policy, in which $\mathbf{W}$ is determined row by row. Other implementation choices, such as prewhitening, the contrast estimation, and the contrast optimization, are discussed in the following subsections, after the introduction of some notations.

4.1. Notation

In practice, source separation problems must be solved with only finite samples of the observed mixtures. In order to account for this limitation, we adopt the following conventions. Vector $\mathbf{x}(t) = [x_1(t), \ldots, x_n(t)]^\mathrm{T}$ denotes the observation of the random vector $\mathbf{x}$ at time t, where t is a discrete time index that belongs to some set of integers $\mathscr{T} \doteq \{1, \ldots, N\}$, with $1 < N < \infty$. The whole sample is denoted by the matrix $\mathbf{X} = [\mathbf{x}(1), \ldots, \mathbf{x}(N)]$, whose rows are written as $\mathbf{x}_i = [x_i(1), \ldots, x_i(N)]$. When referring to the underlying stationary random variable, the time index t is absent. Corresponding notations associated with $\mathbf{s}$ and $\mathbf{y}$ can easily be deduced.

4.2. Whitening

Many successful ICA algorithms such as FastICA and JADE rely on prewhitening because it restricts the space of possible solutions to orthogonal matrices. Another advantage is that the implementation of whitening is straightforward and involves only well studied algebraic procedures, like eigenvalue or singular value decompositions. For instance, if the mixtures are ill-conditioned, this can be detected in the whitening step and this issue can be addressed before the source separation. On the other hand, whitening can also be a source of inaccuracies.

In practice, whitening can be achieved through two different strategies; both of them assume that the mixtures are centered ($\sum_{t=1}^{N} \mathbf{x}(t) = \mathbf{0}$). The simplest strategy relies on the eigenvalue decomposition of the sample covariance of the mixtures, written as

$$\boldsymbol{\Sigma} \doteq \frac{1}{N}\mathbf{X}\mathbf{X}^\mathrm{T} = \boldsymbol{\Xi}\boldsymbol{\Lambda}\boldsymbol{\Xi}^\mathrm{T},$$

where $\boldsymbol{\Xi}$ is an orthogonal matrix and $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues. After the decomposition, the whitening matrix and the whitened mixtures are written as $\mathbf{V} \doteq \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Xi}^\mathrm{T}$ and $\mathbf{Z} = \mathbf{V}\mathbf{X}$, respectively.

We adopt a second strategy that is based on the singular value decomposition of $\mathbf{X}^\mathrm{T}$, denoted

$$\mathbf{X}^\mathrm{T} = \mathbf{U}\mathbf{S}\boldsymbol{\Xi}^\mathrm{T},$$

where $\mathbf{U}$ and $\boldsymbol{\Xi}$ are orthogonal matrices. Matrix $\mathbf{S}$ has the same size as $\mathbf{X}^\mathrm{T}$ and its main diagonal contains the singular values, in descending order; other entries are zero. The decomposition allows us to compute the whitened mixtures directly as the product $\mathbf{Z} = \sqrt{N}\mathbf{I}_{n\times N}\mathbf{U}^\mathrm{T}$, where $\mathbf{I}_{n\times N}$ denotes a truncated $N \times N$ identity matrix, with only n rows. This is a convenient notation to express the fact that we keep only the rows of $\mathbf{U}^\mathrm{T}$ that are associated with the n largest singular values.

The product $\mathbf{Z} = \sqrt{N}\mathbf{I}_{n\times N}\mathbf{U}^\mathrm{T}$ involves only a selection of rows and a scaling. In contrast, the product $\mathbf{Z} = \mathbf{V}\mathbf{X}$ can be slightly less accurate as it involves linear combinations of the rows of $\mathbf{X}$, from a numerical point of view. Additional inaccuracies can stem from the fact that the eigenvalue decomposition in the first strategy does not processes the complete data matrix $\mathbf{X}$. Instead it is applied on the sample covariance matrix $\boldsymbol{\Sigma}$, whose computation can suffer from rounding errors.

In order to be robust against ill-conditioned mixtures, the proposed algorithm checks the results of the singular value decomposition. For this purpose, it estimates the covariance of $\mathbf{Z}$. If some row $\mathbf{z}_i$ has a variance lower than one and/or nonzero covariances with respect to other rows, we remove it from the whitened sample $\mathbf{Z}$; in practice, the algorithm discards all rows of $\mathbf{Z}$ having a standard deviation that is lower than 0.99. Mixtures with a variance lower than one have an abnormally low LAE and thus perturb the subsequent source separation step. As ICA recovers the sources up to a permutation, we can freely rearrange the rows of $\mathbf{Z}$ so that the bad ones are the last p ones; $\mathbf{Z}$ is then updated according to $\mathbf{Z} \leftarrow [\mathbf{z}_1, \ldots, \mathbf{z}_{n-p}]^\mathrm{T}$. Next, the separation algorithm is run on this reduced vector, yielding $n-p$ estimated sources. After the separation, the ICA output consists of the estimated sources, to which we append the badly whitened rows of $\mathbf{Z}$, without any further processing: $\mathbf{Y} \leftarrow [\mathbf{y}_1, \ldots, \mathbf{y}_{n-p}, \mathbf{z}_{n-p+1}, \ldots, \mathbf{z}_n]^\mathrm{T}$. This allows us to measure the separation performance on n signals, as in the normal case.

4.3. Estimation of the contrast

The LAE contrast can be discretized in order to obtain the estimator

$$\hat{\mathscr{C}}_{\mathrm{LAE}}(\mathbf{y}_i) \doteq \min\left\{\left|\min_{t\in\mathscr{T}} y_i(t)\right|, \left|\max_{t\in\mathscr{T}} y_i(t)\right|\right\}. \tag{5}$$

This allows us to reformulate the extraction of a single source with the LAE contrast function as follows: the estimate of the ith source results from the minimization of $\hat{\mathscr{C}}_{\mathrm{LAE}}(\mathbf{w}_i\mathbf{Z})$ with respect to the entries of $\mathbf{w}_i$. Replacing the infimum and supremum of the support with minimum and maximum operators leads to an estimator that is highly sensitive to the sample size, additive noise, and outliers.

Fortunately, better estimators exist and can reliably infer the theoretical infimum and supremum of the support of y_i, starting from the sample vector $\mathbf{y}_i$. As a good tradeoff between evaluation speed and estimation quality is needed, we use averaged order statistics [4,21], as advised in [28] for estimating the range. This estimator works well if the probability distribution function has at least one "sharp frontier". The use of order statistics within the framework of ICA has already been investigated in e.g. [17,1]. Assuming that $\mathbf{y}'_i = [y'_i(1), \ldots, y'_i(N)]$ contains the same entries as $\mathbf{y}_i$ in ascending order, then the estimator based on averaged order statistics can be written as

$$\hat{\mathscr{C}}^{q}_{\mathrm{LAE}}(\mathbf{y}_i) \doteq \min\left\{-\frac{1}{q}\sum_{k=1}^{q} y'_i(k), \frac{1}{q}\sum_{k=1}^{q} y'_i(N+1-k)\right\}, \tag{6}$$

where q is an integer between 1 and $\lfloor N/2 \rfloor$. The estimator is such that

$$\hat{\mathscr{C}}^{q+1}_{\mathrm{LAE}}(\mathbf{y}_i) \leqslant \hat{\mathscr{C}}^{q}_{\mathrm{LAE}}(\mathbf{y}_i). \tag{7}$$

However, if the sample size N is large enough, $(1/q)\sum_{k=1}^{q} y'_i(k) \simeq \min \mathbf{y}_i$ if $y'_i(l) \simeq \min \mathbf{y}_i$ for all $1 \leqslant l \leqslant q$, and likewise for $(1/q)\sum_{k=1}^{q} y'_i(N+1-k)$.

Provided there is no noise and no outliers, we have $\hat{\mathscr{C}}^q_{\mathrm{LAE}}(\mathbf{y}_i) \leqslant \mathscr{C}_{\mathrm{LAE}}(y_i)$ and $\hat{\mathscr{C}}^1_{\mathrm{LAE}}(\mathbf{y}_i) \to \mathscr{C}_{\mathrm{LAE}}(y_i)$ as $N \to \infty$. Hence, the accuracy of the above estimator increases with the sample size N. In the noise-free case, the value of q can be taken close or equal to one. However, it is noteworthy that $\hat{\mathscr{C}}^q_{\mathrm{LAE}}(\mathbf{y}_i)$ is a piecewise linear function and that increasing the value of q “smoothes” it.

In the presence of noise, q must be slightly increased. See [29] for a discussion about the choice of q in the context of range-based contrasts.

The computation of $\hat{\mathscr{C}}_{\mathrm{LAE}}$ involves a sorting operation, whose time complexity ($\mathscr{O}(N \log N)$) is higher than for more traditional contrasts such as the kurtosis (typically $\mathscr{O}(N)$).

4.4. Optimization scheme

Since $\hat{\mathscr{C}}^q_{\mathrm{LAE}}$ is piecewise linear, it is not everywhere differentiable. However, we can reformulate it as a weighted sum over all entries of the ith output sample $\mathbf{y}_i$, which is written as

$$\hat{\mathscr{C}}^q_{\mathrm{LAE}}(\mathbf{y}_i) = \mathbf{y}_i \mathbf{h}_i \tag{8}$$

and where the elements of column vector $\mathbf{h}_i$ are taken in $\{-1/q, 0, 1/q\}$ and depend on the place of $y_i(t)$ in $\mathbf{y}'_i$ and the result of the minimum in (6). If we assume that $\mathbf{h}_i$ is constant in an ε-ball around $\mathbf{w}_i$, then we can compute

$$\mathbf{d} \doteq \left.\frac{\partial \hat{\mathscr{C}}^q_{\mathrm{LAE}}(\mathbf{wZ})}{\partial \mathbf{w}}\right|_{\mathbf{w}_i} = \mathbf{Z}\mathbf{h}_i. \tag{9}$$

The above derivative can be plugged into a basic deflation procedure that is inspired from [14] and works as follows. First, we make $\mathbf{d}$ orthogonal to $\mathbf{w}_i$ and we normalize it:

$$\mathbf{d} \leftarrow \mathbf{d} - \mathbf{d}\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_i, \tag{10}$$

$$\mathbf{d} \leftarrow \frac{\mathbf{d}}{\|\mathbf{d}\|}. \tag{11}$$

Next, we compute a potential new value of the demixing vector $\mathbf{w}$, namely

$$\mathbf{w}'_i \doteq \cos(\alpha)\mathbf{w}_i - \sin(\alpha)\mathbf{d}, \tag{12}$$

where α is an angle lower than $\pi/4$. Notice that $\mathbf{w}'_i$ has a unit norm, just as $\mathbf{w}_i$. Finally, we update $\mathbf{w}_i$ and α, provided the potential new value reduces the LAE. More precisely, if $\hat{\mathscr{C}}^q_{\mathrm{LAE}}(\mathbf{w}'_i\mathbf{z}) < \hat{\mathscr{C}}^q_{\mathrm{LAE}}(\mathbf{w}_i\mathbf{z})$, then α is slightly increased ($\alpha \leftarrow 1.01\alpha$) and we adopt the new value ($\mathbf{w}_i \leftarrow \mathbf{w}'_i$). Otherwise, we leave $\mathbf{w}_i$ unchanged, but we decrease the update angle ($\alpha \leftarrow \alpha/1.2$). After the update, we can re-evaluate the LAE and repeat the above steps in order to converge on a solution.

Notice that the optimization scheme that we propose merely takes into account the direction of $\mathbf{d}$; its norm is neglected. Therefore, the convergence speed mainly depends on the angle α, whose decrease is relatively slow. This choice favors separation quality rather than speed.

Successive updates of $\mathbf{w}_i$ must obviously ensure that $\mathbf{w}_i$ remains orthogonal to all rows of $\mathbf{W}$ that have already been determined by the deflation procedure. Since the mixtures are prewhitened, this means formally that the orthogonality constraint $\mathbf{w}_i\mathbf{w}_j^{\mathrm{T}} = 0$ for $j < i$ should be verified. However, in order to compensate for the accumulation of small inaccuracies in the deflation process, the orthogonality constraint can be relaxed. In practice, we only check that the current row $\mathbf{w}_i$ fulfills the condition $|\mathbf{w}_i\mathbf{w}_j^{\mathrm{T}}| < \cos(\pi/12)$ for all $j < i$. If this is not the case, this means that $\mathbf{w}_i$ converges on some previously found solution and therefore we reinitialize $\mathbf{w}_i$. The new value of $\mathbf{w}_i$ is chosen randomly, so as to be orthogonal to all previous rows.

Table 1

ICALAE: simple deflation procedure to minimize $\hat{\mathscr{C}}^q_{\mathrm{LAE}}$

```
$[\mathbf{W}, \mathbf{V}] = \mathrm{ICALAE}(\mathbf{X})$

(1) Whiten the mixtures using a singular value decomposition:
    (a) Center the sample $\mathbf{X}$ by removing its mean:
        $\mathbf{X} \leftarrow \left[\mathbf{x}(t) - \frac{1}{N}\sum_s^N \mathbf{x}(s)\right]_{1 \leqslant t \leqslant N}$.
    (b) Compute the SVD of the centered sample: $\mathbf{X}^{\mathrm{T}} = \mathbf{USV}^{\mathrm{T}}$.
    (c) Compute $\mathbf{Z}$ directly: $\mathbf{Z} = \sqrt{N}\,\mathbf{I}_{n \times N}\,\mathbf{U}^{\mathrm{T}}$.
(2) Discard the $p$ rows of $\mathbf{Z}$ having a variance lower than $0.99^2$ $(0 \leqslant p \leqslant n)$.
(3) To extract the $i$th source, with $1 \leqslant i \leqslant n - p$, do:
    (a) Initialize $\mathbf{w}_i$ to any random direction and the update angle $\alpha$ to $\pi/4$.
    (b) Check loose orthogonality: if for some $j < i$ the inequality
        $|\mathbf{w}_i\mathbf{w}_j^{\mathrm{T}}| < \cos(\pi/12)$ holds then make $\mathbf{w}_i$ orthogonal to all $\mathbf{w}_j$:
        $\mathbf{w}_i \leftarrow \mathbf{w}_i - \sum_j \mathbf{w}_i\mathbf{w}_j^{\mathrm{T}}\mathbf{w}_j$;   $\mathbf{w}_i \leftarrow \frac{\mathbf{w}_i}{\|\mathbf{w}_i\|}$.
    (c) Compute the $i$th ICA output: $\mathbf{y}_i = \mathbf{w}_i\mathbf{Z}$ (i.e. $y_i(t) = \mathbf{w}_i\mathbf{z}(t)$ for $t \in \mathscr{T}$).
    (d) Estimate the LAE of $\mathbf{y}_i$ using mean order statistics:
        • Determine the indices of the $q$ lowest and $q$ highest values of $y_i(t)$.
        • Average the two corresponding sets of values to obtain the estimated
          infimum and supremum of $y_i(t)$, for $t \in \mathscr{T}$; keep their minimum absolute value
          as in (6) to obtain $\hat{\mathscr{C}}^q_{\mathrm{LAE}}(\mathbf{y}_i)$. Compute vector $\mathbf{h}_i$ accordingly (see (8)).
        • Use $\mathbf{h}_i$ to compute $\mathbf{d} = \mathbf{Z}\mathbf{h}_i$.
        • If $\mathbf{d} \neq \mathbf{w}_i$, make $\mathbf{d}$ orthogonal to $\mathbf{w}_i$ and normalize it:
          $\mathbf{d} \leftarrow \mathbf{d} - \mathbf{d}\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_i$;   $\mathbf{d} \leftarrow \frac{\mathbf{d}}{\|\mathbf{d}\|}$.
    (e) Update $\mathbf{w}_i$ and $\alpha$:
        • Compute $\mathbf{w}'_i = \cos(\alpha)\mathbf{w}_i - \sin(\alpha)\mathbf{d}$ and $\hat{\mathscr{C}}_{\mathrm{LAE}}(\mathbf{w}'_i\mathbf{z})$ (see step (d) above).
        • If $\hat{\mathscr{C}}^q_{\mathrm{LAE}}(\mathbf{w}'_i\mathbf{z}) < \hat{\mathscr{C}}^q_{\mathrm{LAE}}(\mathbf{y}_i)$, then let $\alpha \leftarrow 1.01\alpha$ and $\mathbf{w}_i \leftarrow \mathbf{w}'_i$, else let $\alpha \leftarrow \alpha/1.2$.
    (f) Go back to step 3(b) if convergence is not attained.
(4) Append the $p$ incorrectly whitened mixtures to the extracted sources:
    $\forall i > n - p, \quad \mathbf{y}_i \leftarrow \mathbf{z}_i$.
```

After robust SVD-based whitening, sources are extracted one-by-one with a loose orthogonality constraint preventing error accumulation. The LAE is minimized by a simple gradient descent.

Notice also that a strict orthogonality constraint can be too strong in difficult problems, e.g. with low sample sizes or large numbers of mixtures. In these cases, a small discrepancy between the sample covariances and the true covariances can jeopardize the source extraction after whitening, because whitening decorrelates the sample but not necessarily the underlying random variables. Hence, small departures from orthogonality can allow $\mathbf{W}$ to reach better contrast values and compensate for whitening inaccuracies.

The complete algorithm is detailed in Table 1.

5. The MLSP 2006 data competition

BSS problems were proposed in MLSP 2006 data competition. The statement of the problems conveys some *a priori* information about the sources. For all subproblems, sources are stationary (i.i.d. over time), statistically independent and non-negative. If we assume that u and v are random variables with uniform probability between 0 and 1, two kinds of sources are generated according to

$$s_i = \begin{cases} -\log(u)\max\{0, \mathrm{sign}(v - 0.5)\}, \\ u, \end{cases}$$

where each branch is equiprobable. The first branch yields a super-Gaussian source with sparse distribution, whereas the second possibility leads to a sub-Gaussian source with uniform distribution; see Fig. 4 for an example of both kinds of sources. Therefore, the sample $\mathbf{s}_i$ is a row vector that contains N independent realizations of the corresponding random variable s_i. The mixing matrices are randomly generated, in such a way

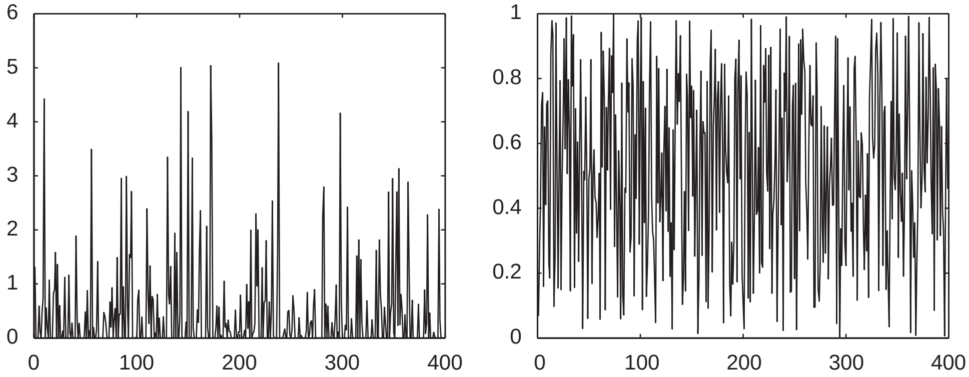

Fig. 4. Sources involved in the MLSP 2006 competition. Sources are statistically independent, stationary and either sparse (left) or uniformly (right) distributed. All sources are non-negative.

that their entries are uniformly distributed between −0.5 and 0.5, unless differently specified. More information regarding the sources and mixing matrices can be found on the competition website (http://mlsp2006.conwiz.dk/).

In order to assess the quality of the source recovery, the competition resorts to the SIR, expressed in dB. The average SIR involves the transfer matrix **C** and can be defined as follows:

$$\mathrm{SIR} = \frac{1}{n}\sum_{i=1}^{n} 10\log_{10}\frac{\max_j c_{ij}^2}{\sum_j c_{ij}^2 - \max_j c_{ij}^2}. \tag{13}$$

The maximum operator is needed because ICA recovers the sources up to a scaling and a permutation. Similarly, it is useful to center and standardize the sources, as most ICA algorithm yields white outputs. Notice that because the SIR is an average, high SIR values do not necessarily mean that *all* sources are well recovered: some sources that are perfectly recovered may compensate for sources that are badly extracted. We expect our algorithm to perform better for sources with a sparse distribution than those with a uniform one. The accuracy of the LAE estimator critically depends on the number of samples near the theoretical endpoint. This number is obviously much lower for the uniform distribution than for the sparse one.

Within the framework of the competition, the average SIR is used in a Monte-Carlo process. The SIR should be higher than 15 dB for at least 90% of the runs, i.e. $P_{90} > 15$ dB, where P_{90} is the 90th percentile. Four sub-problems must be solved:

1. Large scale problem: fixed sample size ($N = 5000$), increasing number of mixtures ($n > 50$), random mixing matrix.
2. Small training set problem: fixed number of mixtures ($n = 50$), decreasing sample size ($N < 5000$), random mixing matrix.
3. Highly ill-conditioned problem: $N = 5000$, $n > 1$; for this subproblem, the mixing matrix is a Hilbert matrix multiplied by a random Givens matrix.
4. Noisy mixtures problem: $n = 50$, $N = 1000$, white noise with increasing variance corrupts the mixtures.

In the above problems, the Hilbert matrix of size n is defined as $\mathbf{H}_n \doteq [1/(i+j-1)]_{1\leqslant i,j\leqslant n}$; its condition number grows as $\mathcal{O}(e^{3.5255n}/\sqrt{n})$. The Givens matrix of size n, denoted by $G_n(p,q,\theta) = [g_{ij}]$, is the same as the identity matrix, except for four elements: $g_{pp} = \cos\theta$, $g_{pq} = \sin\theta$, $g_{qp} = -\sin\theta$, and $g_{qq} = \cos\theta$.

5.1. Large-scale problem

In this first subproblem, the sample size is fixed ($N = 5000$) and the number of sources/mixtures is growing ($n > 50$). The proposed algorithm solves it for a quite large number of mixtures.

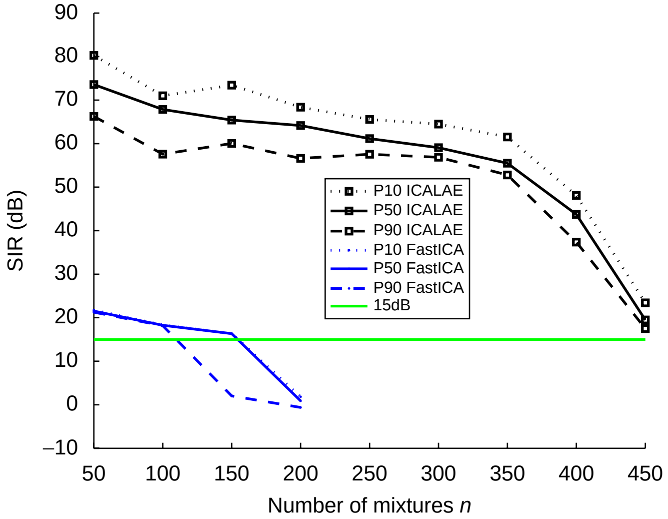

Fig. 5. Results for subproblem 1: SIR performances vs number of sources n for 20 Monte-Carlo runs of ICALAE and FastICA with 5000 sample points. $P_{90} > 15$ dB holds for more than 400 sources.

Graphical results in Fig. 5 shows that outstanding SIR values are attained for more than 400 mixtures (P_{90} is still higher than 30 dB). Processing so many mixtures obviously requires long computation times, even with the fastest algorithms (e.g. FastICA), and justifies the restriction to only 20 Monte-Carlo runs. Average order statistics are computed with $q = 100$. For $N = 5000$ and $n = 50$, the average SIR for sparse sources is around 120 dB. Most uniform sources are recovered with a SIR higher than 25 dB. Approximately 10% of the sources yield a SIR lower than 15 dB. In the same set up, FastICA recovers all sources with SIR values between 20 and 25 dB. If the number of sources increases, ICALAE continues to extract all sparse sources correctly, whereas it progressively fails to extract the uniform ones. Beyond approximately 175 sources, FastICA can no longer extract any source with a SIR higher than 15 dB.

5.2. Small training set problem

In this second problem, the number of sources is kept constant ($n = 50$) but the sample size N varies. The results are shown in Fig. 6 for two algorithms: the proposed one and FastICA (official version 2.5, with “gaus” nonlinearity and fine tuning enabled). Average order statistics are computed with $q = 20$. As can be seen, less than 250 sample points are required to achieve a SIR greater than 15dB in 90% of the cases. In practice, the uniform sources

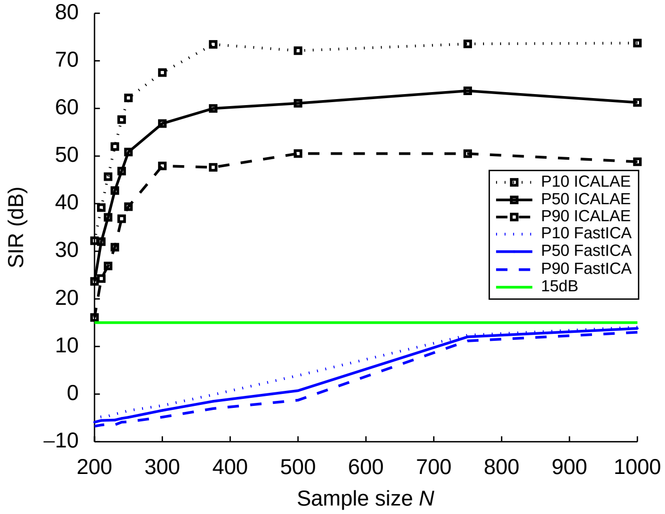

Fig. 6. Results for subproblem 2: SIR performances vs sample size N for 100 Monte-Carlo runs of ICALAE and FastICA with 50 sources. Less than 250 observations are needed to achieve $P_{90} > 15$ dB.

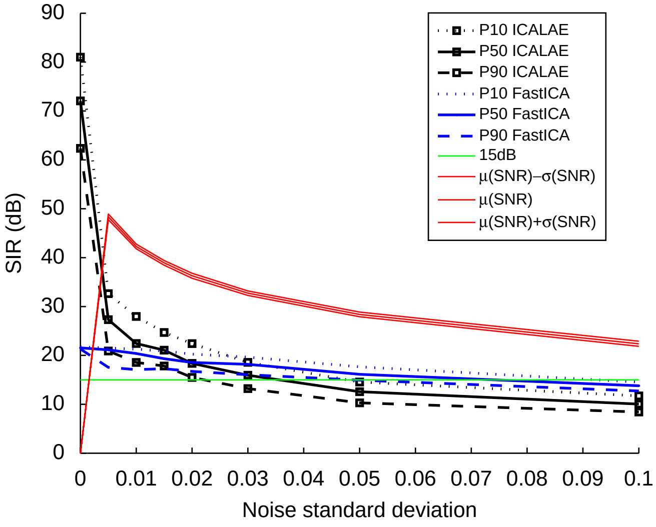

Fig. 8. Results for subproblem 4: SIR performances vs the noise standard deviation, for 100 Monte-Carlo runs of ICALAE and FastICA with 5000 sample points and 50 sources; the corresponding SNR curve is plotted alongside.

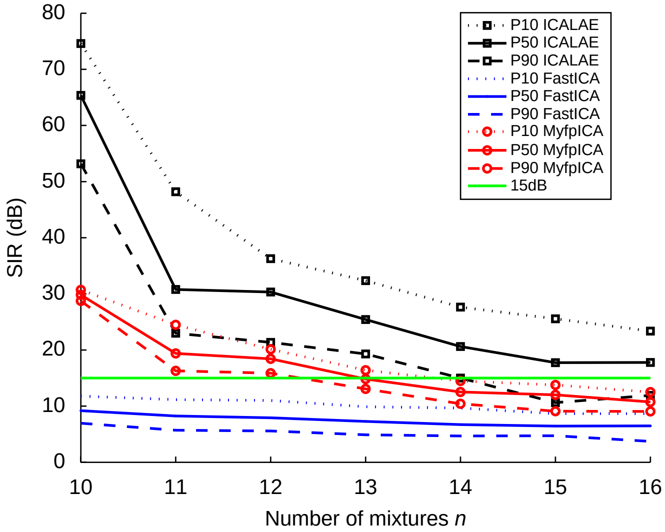

Fig. 7. Results for subproblem 3: SIR performances vs the number of sources n for 100 Monte-Carlo runs of ICALAE ($q = N/200$), FastICA and MyfpICA with 5000 sample points. $P_{90} > 15$ dB holds up to 13 sources.

become more and more difficult to recover as the sample size decreases, whereas a good extraction is still achieved for sparse sources.

5.3. *Highly ill-conditioned problem*

In this third subproblem, the mixing matrix is the product of a Hilbert matrix with a random Givens matrix. Hence, as the number of mixtures is growing, the separation problem gets more and more ill-conditioned. The results are shown in Fig. 7 for three algorithms: the proposed one ($q = 25$), FastICA (as above) and a “hacked” version of FastICA. The latter, called MyfpICA, works with a SVD-based whitening stage and a kurtosis-driven non-linearity (either “kurt” or “gaus” depending on the kurtosis). In this subproblem, achieving a correct whitening is the main difficulty. The proposed algorithm brings a significant performance gain by using the SVD of the centered sample instead of the EVD of the sample covariance matrix. However, beyond 10 mixtures in this problem, the determinant of the mixing matrix **A** is so close to zero that no more than 10 mixtures can be whitened properly, even with the SVD. It has been experimentally observed that additional mixtures after whitening are actually not white; some of them may be correlated and/or have a variance lower than one. In this situation, the trick consists in temporarily discarding these still correlated mixtures after whitening, as proposed in Section 4.4, so that the separation algorithm can run in good conditions.

5.4. *Noisy mixtures problem*

The value of the estimator $\hat{\mathscr{C}}_{\text{LAE}}(y_i)$ relies on a few sample points only, namely on q sample points with $q \ll N$. Consequently the proposed approach is expected not to be very robust against noise and outliers, especially with low values of q; for this problem, q is equal to 100. As can be seen in Fig. 8, the quality of the results is rapidly decreasing as the noise variance is growing. From a more fundamental viewpoint, the literature indicates that the estimation of the distribution endpoint is an intrinsically difficult problem [6].

6. Conclusion

Contrast functions that are based on a range estimation show interesting properties for independent component analysis (ICA) of sources that have a support with finite width. This paper extends this idea to sources whose support is possibly infinite but still bounded on one side. A contrast function that involves an estimator of the least absolute endpoint (LAE) is presented; the LAE is the minimum of the absolute values of the infimum and supremum of the support. The contrast owns the discriminacy property, i.e. all local minima corresponds to solutions of the separation problem.

A specific optimization procedure is proposed; it extracts sources one by one, following a deflation procedure. For each source to be estimated, the LAE is minimized using a simple gradient descent. This allows the proposed algorithm to be quite competitive in terms of speed. The computational cost of a single update is low, as for a fixed-point algorithm, though the latter requires less iterations and converges much faster. The algorithm also relaxes the orthogonality constraint on the separation matrix; it merely checks that all source estimates are distinct. The main advantage of this milder constraint is that a deflation approach can be used without accumulating errors in the successive source estimates. Finally, the optimization of the whitening step also improves the performances of the algorithm, especially for ill-conditioned mixtures.

The proposed contrast and algorithm are shown to perform efficiently on at least three ICA problems from the IEEE MLSP 2006 competition presented by A. Cichocki and D. Erdogmus.

References

[1] Y. Blanco, S. Zazo, An overview of BSS techniques based on order statistics: formulation and implementation issues, in: C. Puntonet, A. Prieto (Eds.), Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation (ICA), Lecture Notes in Computer Science, vol. 3195, Springer, Granada (Spain), 2004, pp. 73–80.

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

[3] S. Cruces, I. Duran, The minimum support criterion for blind source extraction: a limiting case of the strengthened Young's inequality, in: C. Puntonet, A. Prieto (Eds.), Proceedings of the International Conference Independent Component Analysis and Blind Signal Separation (ICA 2004), Lecture Notes in Computer Science, vol. 3195, Springer, Granada (Spain), 2004, pp. 57–64.

[4] H. David, Order Statistics, Wiley Series in Probability and Mathematical Statistics, Wiley, New York, 1970.

[5] A. Erdogan, A simple geometric blind source separation method for bounded magnitude sources, IEEE Trans. Signal Process. 54 (2) (2006) 438–449.

[6] A. Goldenshluger, A. Tsybakov, Estimating the endpoint of a distribution in the presence of additive observation errors, Stat. Probab. Lett. 68 (2004) 39–49.

[7] P. Gruber, K. Stadlthanner, M. Bohm, F.J. Theis, E.W. Lang, A. Tome, A.R. Teixeira, C.G. Puntonet, J.M. Gorriz Saez, Denoising using local projective subspace methods, Neurocomputing 69 (13–15) (2006) 1485–1501 Blind Source Separation and Independent Component Analysis—Selected papers from the ICA 2004 meeting, Granada, Spain, Blind Source Separation and Independent Component Analysis.

[8] P. Hall, On estimating the endpoint of a distribution, Ann. Stat. 10 (1982) 556–568.

[9] S. Haykin, Z. Chen, The cocktail party problem, Neural Comput. 17 (2005) 1875–1902.

[10] A. Hyvärinen, Fast and robust fixed-point algorithms for independent component analysis, IEEE Trans. Neural Networks 10 (3) (1999) 626–634.

[11] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley Series on Adaptive Learning Systems for Signal Processing, Communications, and Control, Wiley, New York, 2001.

[12] A. Hyvärinen, E. Oja, A fast fixed-point algorithm independent component analysis, Neural Comput. 9 (7) (1997) 1483–1492.

[13] C.J. James, C.W. Hesse, Independent component analysis for biomedical signals, Physiol. Meas. 26 (2004) R15–R39.

[14] J. Lee, F. Vrins, M. Verleysen, A simple ICA algorithm for non-differentiable contrasts, in: Proceedings of the European Signal Processing Conference (EUSIPCO), Antalya, Turkey, 2005, pp. cr1412.1–4.

[15] J. Lee, F. Vrins, M. Verleysen, Non-orthogonal support width ICA, in: M. Verleysen (Ed.), Proceedings of the European Symposium on Artificial Neural Networks (ESANN), Bruges, Belgium, 2006, pp. 351–358.

[16] A. Mansour, C. Jutten, What should we say about the kurtosis, Signal Process. Lett. 6 (12) (1999) 321–322.

[17] D.-T. Pham, Blind separation of instantaneous mixtures of sources based on order statistics, IEEE Trans. Signal Process. 48 (2) (2000) 363–375.

[18] D.-T. Pham, Contrast functions for blind separation and deconvolution of sources, in: T. Lee, T. Jung, S. Makeig, T. Sejnowski (Eds.), Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation (ICA), San Diego, CA, 2001, pp. 37–42.

[19] D.-T. Pham, Mutual information approach to blind separation of stationary sources, IEEE Trans. Inf. Theory 48 (7) (2002) 1935–1946.

[20] M. Plumbley, Algorithms for nonnegative independent component analysis, IEEE Trans. Neural Networks 4 (3) (2003) 534–543.

[21] C. Rose, M. Smith, Order Statistics, Mathematical Statistics with Mathematica, Springer, New York, 2002.

[22] F. Vrins, Contrast properties of entropic criteria for blind source separation: a unifying framework based on information-theoretic inequalities, Ph.D. Thesis, Université catholique de Louvain, Louvain-la-Neuve (Belgium) (March 2007). URL ⟨http://edoc.bib.ucl.ac.be:81/ETD-db/collection/available/BelnUcetd-02162007-112342/unrestricted/PhD_Vrins_Thesis.pdf⟩.

[23] F. Vrins, C. Jutten, M. Verleysen, SWM: a class of convex contrasts for source separation, in: Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP), IEEE, Philadelphia, PA, 2005, pp. V.161–V.164.

[24] F. Vrins, J. Lee, M. Verleysen, Filtering-free blind separation of correlated images, in: A.P.J. Cabestany, F. Sandoval (Eds.), Proceedings of the International Work-conference on Artificial Neural Networks (IWANN): Computational Intelligence and Bioinspired Systems, Lecture Notes in Computer Science, vol. 3512, Springer, 2005, pp. 1091–1099.

[25] F. Vrins, J.A. Lee, M. Verleysen, A minimum-range approach to blind extraction of bounded sources, IEEE Trans. Neural Networks 18 (3) (2007) 809–822.

[26] F. Vrins, D.-T. Pham, Minimum range approach to blind partial simultaneous separation of bounded sources: contrast and discriminacy properties, Neurocomputing 70 (7–9) (2007) 1207–1214.

[27] F. Vrins, D.-T. Pham, M. Verleysen, Mixing and non-mixing local minima of the entropy contrast for blind source separation, IEEE Trans. Inf. Theory 3 (53) (2007) 1030–1042.

[28] F. Vrins, M. Verleysen, Minimum support ICA using order statistics. Part I: quasi-range based support estimation, in: J. Rosca, et al. (Ed.), Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2006), Lecture Notes in Computer Science, vol. 3889, Springer, Charleston, SC, 2006, pp. 262–269.

[29] F. Vrins, M. Verleysen, Minimum support ICA using order statistics. Part II: performance analysis, in: J. Rosca, et al. (Ed.), Proceedings of the International Conference Independent Component Analysis and Blind Signal Separation (ICA 2006), Lecture Notes in Computer Science, vol. 3889, Springer, Charleston, SC, 2006, pp. 270–277.

[30] H. Yang, S. Amari, Adaptive on-line learning algorithms for blind separation—maximum entropy and minimum mutual information, Neural Comput. 9 (1997) 1457–1482.

[31] P. Zibulevsky, B.A. Pearlmutter, Blind source separation by sparse decomposition, Neural Comput. 13 (4) (2001) 863–882.

John Aldo Lee was born in 1976 in Brussels, Belgium. 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, both from the Université catholique de Louvain (UCL, 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 F.N.R.S. (Fonds National de la Recherche Scientifique). His current work aims at developing specific image enhancement techniques for Positron Emission Tomography in the Molecular Imaging and Experimental Radiotherapy department of the Saint-Luc University Hospital (Belgium).

Frédéric Vrins was born in Uccle, Belgium, in 1979. He received his M.Sc. degree in mechatronics engineering (2002), his D.E.A. degree in Applied Sciences (2004) and his Ph.D. degree (2007) from the Université catholique de Louvain (Belgium). His research interests are blind source separation, independent component analysis, Shannon and Rényi entropies, mutual information and information theory in adaptive signal processing. Frédéric Vrins is currently working as a quantitative analyst (modeler) within the Financial Markets Department of ING Wholesale Banking (Brussels, Belgium).

Michel Verleysen was born in 1965 in Belgium. He received his M.Sc. and Ph.D. degrees in electrical engineering from the Université catholique de Louvain (Belgium) in 1987 and 1992, respectively. He was an invited professor at the Swiss E.P.F.L. (Ecole Polytechnique Fédérale de Lausanne, Switzerland) in 1992, at the Université d'Evry Val d'Essonne (France) in 2001, and at the Université ParisI-Panthéon- Sorbonne from 2002 to 2007, respectively. He is now a professor at the Universite' catholique de Louvain, and Honorary Research Director of the Belgian F.N.R.S. (National Fund for Scientific Research). He is editor-in-chief of the Neural Processing Letters journal, chairman of the annual ESANN conference (European Symposium on Artificial Neural Networks), associate editor of the IEEE Transactions on Neural Networks journal, and member of the editorial board and program committee of several journals and conferences on neural networks and learning. He is author or co-author of more than 200 scientific papers in international journals and books or communications to conferences with reviewing committee. He is the co-author of the scientific popularization book on artificial neural networks in the series "Que Sais-Je?" in French, and of the "Nonlinear Dimensionality Reduction" book published by Springer in 2007. His research interests include machine learning, artificial neural networks, self-organization, time-series forecasting, nonlinear statistics, adaptive signal processing, and high-dimensional data analysis.