Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®
Signal Processing 85 (2005) 1029–1044
SIGNAL PROCESSING
www.elsevier.com/locate/sigpro

On the entropy minimization of a linear mixture of variables for source separation

Frédéric Vrins*, Michel Verleysen[1]

Université Catholique de Louvain - UCL, Electrical Engineering Department, Microelectronics Laboratory, Machine Learning Group, Maxwell Building, Place du Levant 3, 1348 Louvain-la-Neuve, Belgium

Received 26 April 2004; received in revised form 29 September 2004

Abstract

The marginal entropy $h(Z)$ of a weighted sum of two variables $Z = \alpha X + \beta Y$, expressed as a function of its weights, is a usual cost function for blind source separation (BSS), and more precisely for independent component analysis (ICA). Even if some theoretical investigations were done about the relevance from the BSS point of view of the global minimum of $h(Z)$, very little is known about possible local spurious minima.

In order to analyze the global shape of this entropy as a function of the weights, its analytical expression is derived in the ideal case of independent variables. Because of the ICA assumption that distributions are unknown, simulation results are used to show how and when local spurious minima may appear. Firstly, the entropy of a whitened mixture, as a function of the weights and under the constraint of independence between the source variables, is shown to have only relevant minima for ICA if at most one of the source distributions is multimodal. Secondly, it is shown that if independent multimodal sources are involved in the mixture, spurious local minima may appear. Arguments are given to explain the existence of spurious minima of $h(Z)$ in the case of multimodal sources. The presented justification can also explain the *location* of these minima knowing the source distributions. Finally, it results from numerical examples that the maximum-entropy mixture is not necessarily reached for the 'most mixed' one (i.e. equal mixture weights), but depends of the entropy of the mixed variables.

Keywords: Entropy; Local minima; Blind source separation; ICA

1. Introduction

The marginal entropy of a whitened weighted sum of m variables is an important measure in many data and signal processing contexts. For instance, the signal entropy can be used as a cost

*Corresponding author.
E-mail addresses: vrins@dice.ucl.ac.be (F. Vrins), verleysen@dice.ucl.ac.be (M. Verleysen).
[1]M.V. is a Senior Research Associate with the Belgian National Fund for the Scientific Research (FNRS).

0165-1684/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2004.11.026

function for blind source separation (BSS) (see e.g. [10]) and blind deconvolution [14,15], in the independent component analysis (ICA) context. Indeed, in such applications, efficient algorithms minimizing this entropic function can lead to recovering the lowest entropic source [10]. This method can be extended to the simultaneous extraction of $n \leqslant m$ sources if n non-singular mixtures are available.

Despite its popularity in the ICA community, very little is known about the global shape of the marginal entropy, when it is expressed as a function of the mixing coefficients. However, having information about this shape can be critical when the minimization of the marginal entropy function is performed through a gradient descent as it is the case in most ICA algorithms. Actually, some spurious local minima of the entropy may appear, as mentioned in [3]. In [16], Cardoso presents a simple example involving two independent sources having a joint distribution with $3 \times 3 = 9$ modes; in this example the maximum likelihood-based cost function, another criterion for ICA, fails to give a satisfactory non-mixing solution, even when the source distributions are a priori known. An intuitive justification can explain this problem, looking at the Kullback–Leibler distance between the target and the output distributions. In this paper, we extend this study to the marginal entropy cost function (equivalent, under the whiteness constraint, to the mutual information one [7]).

This work covers the problem of a mixture of two sources. The influence of the source independence assumption and of the source distribution modality on the existence of spurious minima is analyzed through numerical simulations. On one hand, simulation results show that the dependence level between the original sources influences the existence of spurious minima. Furthermore, it is shown that the most random mixture is not necessarily obtained with equal mixture coefficients: the level of *randomness* of the mixture also depends on the source entropies. On the other hand, the impact of the source distributions shape is examined; two simple examples involving two independent sources having a joint distribution with $2 \times 2 = 4$ modes are given; in this examples, the marginal entropy cost function fails to give an acceptable, non-mixing, solution. An intuitive explanation of this phenomenon is provided in terms of structural modification (actually, the number of modes) of the distribution. The simulation results are reinforced by theoretical arguments (without using any information on the source distributions, except their (in)dependence) when possible.

This paper is organized as follows. In Section 2, the problem of source separation is briefly recalled, and the impact of whitening the signals issued from ICA on the mixing coefficients is derived. The ICA application is presented as an optimization problem in Section 3. In Section 4, the entropy is defined, and its use for ICA is argued from two point of views. The output distributions are characterized in Section 5. In Sections 6 and 7, the impact of source independence and source modality is analyzed, respectively.

2. Blind source separation

2.1. Separation consists in adjusting the weights of a mixture

Consider a vector of n unknown zero-mean sources $\mathbf{S} = [S_1, \ldots, S_n]^{\mathrm{T}}$ linearly combined by an unknown square $n \times n$ (invertible) mixing matrix $\mathbf{A}$, resulting in a vector of observed signals $\mathbf{U} = [U_1, \ldots, U_n]^{\mathrm{T}}$ (the T exponent denotes the transposition operator):

$$\mathbf{U} = \mathbf{AS}. \tag{1}$$

The sources extraction method consists in multiplying $\mathbf{U}$ by an unmixing matrix $\mathbf{B}$ such that $\mathbf{W} = \mathbf{BA}$ is diagonal, up to permutations between its rows (one non-zero element per row and per column: $\mathbf{W}$ is 'non-mixing'). The vector of the outputs $\mathbf{Z} \doteq \mathbf{BU} = \mathbf{WS}$ is thus an estimation of the original sources, possibly permuted and scaled.

Now assume that $\mathbf{S} = [S_1\, S_2]^{\mathrm{T}}$ $(n = 2)$ and that Z is the first element of $\mathbf{Z}$ $(Z \doteq Z_1)$. In this case, the mixing system reduces to

$$\mathbf{U} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} \tag{2}$$

and

$$Z = \underbrace{(b_{11}a_{11} + b_{12}a_{21})}_{\alpha} S_1 + \underbrace{(b_{11}a_{12} + b_{12}a_{22})}_{\beta} S_2, \tag{3}$$

where a_{ij} and b_{ij} are the elements of $\mathbf{A}$ and $\mathbf{B}$, respectively. The invertibility of the mixing system implies that the rank of $\mathbf{A}$ must be two ($\det|\mathbf{A}| \neq 0$). Consequently, by adjusting the elements b_{1i} of the first row of $\mathbf{B}$ (or equivalently the elements α and β of the first row of $\mathbf{W}$), it is possible to yield Z proportional to an original source. In the following, we consider that α and β (called here 'weights' or 'mixture coefficients') can be freely adjusted, without loss of generality.

It is well known (see e.g. [17]) that if the extraction of both sources is required, Z_2 can be obtained by orthogonalization of $\mathbf{B}$ subject to $\mathbf{\Sigma_Z} = \mathbf{I_B}$, where $\mathbf{\Sigma_Z}$ is the covariance matrix of $\mathbf{Z}$ and $\mathbf{I_B}$ the identity matrix of the same size as $\mathbf{B}$. This is called a *deflation* approach [13].

2.2. Whitening all signals

If the zero-mean signals $\mathbf{Z}$ have a covariance matrix equal to identity they are said to be *white*: $\mathbf{\Sigma_Z} = \mathbf{I_B}$ [17]. Under whitening constraint on the output Z, the coefficients α and β are not independent. Indeed, constraining the variance $\sigma_Z^2 = 1$ we have

$$\begin{aligned}
\sigma_Z^2 &\doteq E\{(\alpha S_1 + \beta S_2)^2\} - E\{\alpha S_1 + \beta S_2\}^2 \\
&= \alpha^2 E\{S_1^2\} + \beta^2 E\{S_2^2\} + 2\alpha\beta E\{S_1 S_2\} - E\{\alpha S_1\}^2 \\
&\quad - E\{\beta S_2\}^2 - 2\alpha\beta E\{S_1\}E\{S_2\} \\
&= \alpha^2 \underbrace{(E\{S_1^2\} - E\{S_1\}^2)}_{\sigma_{S_1}^2} + \beta^2 \underbrace{(E\{S_2^2\} - E\{S_2\}^2)}_{\sigma_{S_2}^2} \\
&\quad + 2\alpha\beta \underbrace{(E\{S_1 S_2\} - E\{S_1\}E\{S_2\})}_{\delta_{S_1S_2}}.
\end{aligned} \tag{4}$$

Moreover, as the sources are independent, they are uncorrelated ($\delta_{S_1S_2} = 0$) and the previous relation reduces to

$$\sigma_Z^2 = \alpha^2 + \beta^2 = 1. \tag{5}$$

The α and β coefficients may then be expressed as circular functions of an angle θ,

$$\begin{aligned}
\alpha &\doteq \sin(\theta), \\
\beta &\doteq \cos(\theta).
\end{aligned} \tag{6}$$

The observed signals are also supposed to be uncorrelated and to have unit variance (i.e. they are whitened: $\mathbf{\Sigma_U} = \mathbf{I_B}$). If they are not, a linear transformation based on the eigenvalue decomposition of $\mathbf{\Sigma_U}$ is applied [8]. It has been shown in [17] that this constraint combined with $\mathbf{\Sigma_Z} = \mathbf{I_B}$ reduces $\mathbf{B}$ to an orthogonal matrix ($|\det(\mathbf{B})| = 1$).

Note that the whiteness assumption on the sources is not restrictive if the sources are uncorrelated. Indeed, we may divide S_i by σ_{S_i} and multiply the ith column of $\mathbf{A}$ by the same quantity, without changing the observed signals $\mathbf{U}$.

In the remaining of this paper, the sources will be deemed to have a unitary variance. In some examples, they will be slightly correlated, but in that cases the value of the Pearson's coefficient will be given. In all situations, we will impose the constraint on the weights given by Eq. (5), without guarantee that it will imply $\sigma_Z^2 = 1$, since $\delta_{S_1S_2}$ can be non-zero.

3. From BSS to ICA and optimization

Of course, in 'real-world' applications *blind* update rules for the elements of $\mathbf{B}$ must be found; *blind* means that neither the sources nor the mixing matrix are known, so that the trivial solution $\mathbf{B}^\star \doteq \mathbf{A}^{-1}$ cannot be used.

If the sources $\mathbf{S}$ are mutually independent, the ICA methods allow to recover the original signals by minimizing a cost function $\mathscr{C}$; the main property of adequate cost functions is that $\mathscr{C}(\mathbf{Z}) \geqslant 0$, with equality (global minimum) if and only if the output signals are independent, i.e. if

$$f_{\mathbf{Z}} = f_{Z_1} f_{Z_2}, \tag{7}$$

where $f_{\mathbf{Z}}$ and f_{Z_i} are the joint and marginal distributions of $\mathbf{Z}$ and Z_i respectively. If at most

one source S_i has a Gaussian distribution, it is known from the Darmois–Skitovitch theorem [8,12] that these independent components correspond to the original sources:

$$\mathbf{Z}^{\diamond} = \underbrace{\mathbf{B}^{\diamond}\mathbf{A}}_{\mathbf{W}^{\diamond}}\mathbf{S} = \mathbf{PDS}, \tag{8}$$

where $\mathbf{P}$ and $\mathbf{D}$ denote permutation and diagonal matrices, respectively. Similarly, $\mathbf{Z}^{\diamond}$, $\mathbf{B}^{\diamond}$ and $\mathbf{W}^{\diamond}$ are the $\mathbf{Z}$, $\mathbf{B}$ and $\mathbf{W}$ matrices found at $\mathscr{C}(\mathbf{Z}) = 0$. In other words, $\mathbf{W}^{\diamond}$ is non-mixing. In the remaining of this paper, it is assumed that at most one source follows the Normal law.

Several cost functions $\mathscr{C}$ were derived for ICA, for example using the mutual information [16,17,8], the kurtosis [16,17,6], the order statistics [19], the higher-order cumulants [10,16,17,5] or the marginal entropy of one of the outputs (see a.o. [3,11]). In this paper, we focus on this latter criterion.

4. Differential entropy as cost function for ICA

The *minimum marginal entropy* is one of the cost functions that allow recovering independent signals. If the original sources $\mathbf{S}$ are mutually independent, finding the unmixing matrix $\mathbf{B}$ that minimizes—globally—the entropy $h(Z)$ of the mixture Z will imply $\beta = 0$ if S_1 has a lower entropy than S_2, and $\alpha = 0$ in the opposite case. The other signal can be found by deflation (see Section 2.1). An alternative to the deflation approach allowing to extract simultaneously all the sources is to minimize the sum of the output marginal entropies: $h(Z_1) + h(Z_2)$.

In this section, we first introduce the differential entropy $h(X)$ of a random variable X. Next, the relevance of $h(Z_1) + h(Z_2)$ and $h(Z)$ as cost functions for ICA is examined starting from the mutual information (MI) concept and using the entropy power inequality (EPI).

4.1. Differential entropy

Definition 1. The differential entropy (also called Shannon's entropy) of a continuous variable X with distribution f_X is defined as [21]

$$h(X) \doteq -\int f_X(x)\log f_X(x)\,\mathrm{d}x. \tag{9}$$

This definition applies both to univariate and multivariate random variables. The differential entropy $h(X)$ can be seen as a measure of the amount of 'randomness' in the variable X. In this paper, the theoretical developments are presented with logarithm to base 2 for simplicity, but the results remain valid if other bases are chosen. The following theorem illustrates a property of the differential entropy that will be used further in this paper.

Theorem 1. *Consider $X^{\star} = CX + \mu$, where μ is a shift, C a scaling (matrix or scalar) factor and X a random variable. Then, the entropy of $X^{\star}$ is given by* [9]

$$h(X^{\star}) = h(X) + \log|C|. \tag{10}$$

If C is a matrix, $|C|$ denotes the absolute value of $\det(C)$. This expression shows that the entropy is not sensitive to shifting but well to scaling. Note that contrarily to the entropy of a discrete variable, the differential entropy can be negative. Furthermore, if C is an orthogonal matrix, $h(X^{\star}) = h(X)$.

4.2. Minimum mutual information approach

The mutual information (MI)—noted I—between two variables Z_1 and Z_2 is the Kullback–Leibler divergence (also called *relative entropy*) between the joint density and the product of the marginal ones [9]:

$$I(\mathbf{Z}) = \int f_{\mathbf{Z}}(z)\log\frac{f_{\mathbf{Z}}(z)}{f_{Z_1}(z_1)f_{Z_2}(z_2)}\,\mathrm{d}z. \tag{11}$$

The mutual information can be used as a measure of independence, and was proposed as a cost function for ICA by Comon [8]. It can be shown [9] that the MI can be rewritten in terms of entropies:

$$I(\mathbf{Z}) = h(Z_1) + h(Z_2) - h(\mathbf{Z}). \tag{12}$$

As previously explained in Section 2.2, $\mathbf{B}$ is orthogonal under whitening constraint. Therefore

Eq. (10) gives $h(\mathbf{Z}) = h(\mathbf{U}) = h(\mathbf{S}) + \log|\det(\mathbf{A})|$: the joint entropy of the whitened outputs is constant under invertible transformations [7,9] (since $|\det \mathbf{B}| = 1$). In the case of whitened outputs, minimizing the mutual information (i.e. the dependence) between the outputs is equivalent to minimizing the sum of the *marginal entropies*, since the last term of the right-hand part of Eq. (12) is constant. Note that the *marginal entropy* criterion is different from the one used in the algorithm derived by Bell and Sejnowski [2]. Indeed, this algorithm tries to maximize the *joint entropy* of the outputs, non-linearly mapped in $[0,1]^n$ through an activation function. Hence, the sum of the marginal output entropies is a local cost function for BSS:

$$\mathscr{C}_{12}(\mathbf{Z}) \doteq h(Z_1) + h(Z_2). \tag{13}$$

4.3. Entropy power inequality approach

A fundamental result in statistics and information theory is the entropy power inequality.

Theorem 2. *If S_1 and S_2 are two independent random variables, then* [9]:

$$2^{2h(S_1+S_2)} \geqslant 2^{2h(S_1)} + 2^{2h(S_2)} \tag{14}$$

with equality if and only if S_1 and S_2 follow the Normal law.

The entropy power inequality is used to prove the following property of the entropy of a weighted sum of random variables.

Corollary 1.

$$h(Z) = h(\alpha S_1 + \beta S_2) \geqslant \max(h(S_1) + \log|\alpha|,\, h(S_2) + \log|\beta|). \tag{15}$$

Proof. By substitution, we have that $2^{2h(\alpha S_1+\beta S_2)} \geqslant 2^{2h(\alpha S_1)} + 2^{2h(\beta S_2)}$. Using Eq. (10), the fact that $2^{2h(\Upsilon)} \geqslant 0$ for any random variable Υ, and the strictly increasing property of the logarithm function, we have the enounced result. □

The corresponding 'forbidden zones' for $h(Z)$ are colored in dark in Fig. 1. For illustration purposes, $h(S_1)$ was arbitrarily chosen lower than $h(S_2)$.

Unfortunately, the result enounced by Eq. (15) does not imply that $h(\alpha S_1 + \beta S_2) \geqslant \min(h(S_1), h(S_2))$ (see small white area below the horizontal solid line in Fig. 1). Nevertheless, this last relation can easily be proven as shown below.

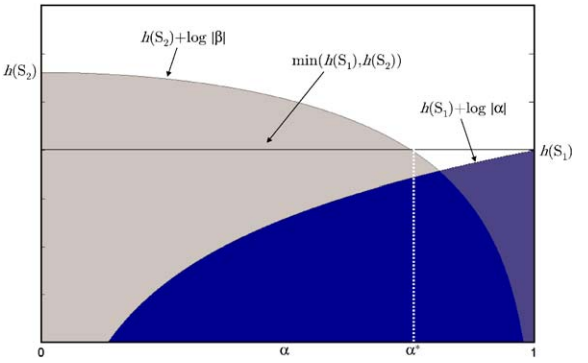

Fig. 1. Restricted areas in the entropy space. The marginal entropy of a whitened weighted mixture is located above the zone delimited by the dark areas (Eq. (15)) and the horizontal line $\Pi \equiv \min(h(S_1), h(S_2))$ (Eq. (16)).

Corollary 2. *If S_1 and S_2 are independent random variables and if at least one of them has not a Gaussian distribution, then the following inequality holds*:

$$h(Z) = h(\alpha S_1 + \beta S_2) \geqslant \min(h(S_1),\, h(S_2)) \tag{16}$$

with equality if and only if $\alpha = 0$ in the $h(S_2) < h(S_1)$ case, $\beta = 0$ in the $h(S_1) < h(S_2)$ case or if $\alpha\beta = 0$ in the $h(S_2) = h(S_1)$ case.

Proof. In the first part of the proof, it is assumed that $\alpha\beta \neq 0$. Hence, Theorem 2 can be applied on αS_1 and βS_2 instead of S_1 and S_2, respectively. Using the hypothesis that both source distributions are non-Gaussian, then, combining Theorems 1 and 2, we have

$$2^{2h(\alpha S_1+\beta S_2)} > 2^{2h(\alpha S_1)} + 2^{2h(\beta S_2)} \tag{17}$$

$$= 2^{2(h(S_1)+\log|\alpha|)} + 2^{2(h(S_2)+\log|\beta|)} \tag{18}$$

$$= 2^{2h(S_1)} 2^{\log|\alpha^2|} + 2^{2h(S_2)} 2^{\log|\beta^2|} \tag{19}$$

$$= \alpha^2 2^{2h(S_1)} + \beta^2 2^{2h(S_2)}. \tag{20}$$

Since the logarithm is a strictly increasing function, we can rewrite this expression as

$$h(\alpha S_1 + \beta S_2) > \tfrac{1}{2}\log(\alpha^2 2^{2h(S_1)} + \beta^2 2^{2h(S_2)}). \tag{21}$$

Now, assume that $h(S_2) \geqslant h(S_1)$. Taking into account that the outputs are whitened (Eq. (5) holds), we obtain the following result:

$$h(\alpha S_1 + \beta S_2) > \tfrac{1}{2}\log(2^{2h(S_1)} + \underbrace{\beta^2(2^{2h(S_2)} - 2^{2h(S_1)})}_{\geqslant 0}) \tag{22}$$

$$> h(S_1). \tag{23}$$

Hence, we have $h(\alpha S_1 + \beta S_2) > h(S_1)$. On the contrary, if $h(S1) \geqslant h(S2)$, it is proven in the same way that $h(\alpha S1 + \beta S2) > h(S2)$. In other words, if $\alpha\beta \neq 0$, the equality in Eq. (16) cannot be reached: $h(\alpha S1 + \beta S2) > \min(h(S_1), h(S_2))$.

The specific cases of $\alpha = 0$ and/or $\beta = 0$ are handled differently. It is obvious that if $\alpha = 0$, then $Z = S_2$ implying $h(Z) = h(S_2) \geqslant \min(h(S_1), h(S_2))$; the equality is reached if and only if $h(S_2) \leqslant h(S_1)$. Similarly, if $\beta = 0$, then $Z = S_1$ and $h(Z) = h(S_1) \geqslant \min(h(S_1), h(S_2))$; the equality holds if and only if $h(S_1) \leqslant h(S_2)$. Furthermore, if $h(S_1) = h(S_2)$, then both $\alpha = 0$ or $\beta = 0$ ensure that $h(Z) = \min(h(S_1), h(S_2))$. The combination of these results proves Eq. (16). □

Corollary 2 can be observed in Fig. 1.

As a consequence, the entropy of a unit-variance weighted mixture of two whitened independent sources is larger or equal than the entropy of the lowest entropic source. It follows from Eq. (23) that the equality is reached when the absolute value of the weight associated to this particular source is one (all the other weights being zero); the entropy of a marginal output distribution is a local cost function for ICA:

$$\mathscr{C}_i(Z_i) \doteq h(Z_i). \tag{24}$$

Note that contrarily to the MI approach, the entropy of a marginal distribution focus on a single output signal $Z = Z_i$.

4.4. Gradient descent on $\mathscr{C}_{12}(\mathbf{Z})$ and $\mathscr{C}_i(Z_i)$?

The above sections show that minimizing the sum of the output entropies or the entropy of a single output allows one to solve the BSS problem. Furthermore, the global minimum of $\mathscr{C}_i$ leads to recover the lowest entropic source (up to a scale factor). A generalization to more than two sources may be found in [11]; however in the latter, Poincaré's separation theorem of matrix algebra must be used.

If one wants to find *non-mixing* minima (corresponding to $\mathbf{B} = \mathbf{B}^{\diamond}$) by gradient descent, the existence of *mixing* local minima (i.e. minima that do not correspond to $\{\alpha\beta = 0$ and $\alpha + \beta \neq 0\}$ and thus associated to spurious solutions) of $h(Z)$ must be discussed; this is the topic of the next sections. Note that if mixing minima exist, they are local minima because of Corollary 2 (see Eq. (16)).

5. Characterization of the output distributions

The analytical expression of $h(Z)$ expressed as a function of the weights and of the differential entropy of the sources will be derived. It is well known (see e.g. [20]) that if Z is a sum of independent variables S_1 and S_2, then f_Z is the convolution of the distributions f_{S_1} and f_{S_2}. For instance,

$$f_Z = f_{S_1} * f_{S_2} \quad \text{or} \quad f_Z(z) = \int f_{S_1}(\tau) f_{S_2}(z - \tau)\,\mathrm{d}\tau, \tag{25}$$

where the symbol $*$ denotes the convolution operator. Therefore, the distribution of $Z = \alpha S_1 + \beta S_2$ is the convolution of $f_{\alpha S_1}$ and $f_{\beta S_2}$:

$$f_Z = f_{\alpha S_1} * f_{\beta S_2}. \tag{26}$$

A change of variables makes it possible to rewrite the densities $f_{\alpha S_1}$ and $f_{\beta S_2}$ in terms of the marginal distributions of the original variables. Indeed, if $V = \alpha S_1$, then [9]

$$f_V(v) = f_{\alpha S_1}(v) = \frac{f_{S_1}(v/\alpha)}{|\alpha|}. \tag{27}$$

Using Eqs. (26) and (27), the distribution of Z can be rewritten as

$$f_Z(z) = \frac{1}{|\alpha\beta|} \int f_{S_1}\left(\frac{\tau}{\alpha}\right) f_{S_2}\left(\frac{z - \tau}{\beta}\right) \mathrm{d}\tau. \tag{28}$$

This analytical expression does not allow to draw conclusions on the existence of mixing minima of $h(Z)$ without knowing some information about the densities of the sources S_1 and S_2 involved in the mixture. In particular, the structure of the

distribution (unimodal or multimodal) influences the existence of mixing minima, even in the ideal mixing case (see Section 7). For this reason, in order to discuss the significance of each local minimum from the source separation point of view (and to extend the justification to dependent sources), the next sections present simulation results of the entropy of a mixture.

Let us remark that $\mathscr{C}_i(Z_i)$ cannot be used as cost function for BSS in the case of Gaussian sources. Indeed, having Eq. (26) in mind, it can be proven that if both f_{S_1} and f_{S_2} are Normal distributions, then f_Z is also a Normal distribution. Moreover, the variance of Z is unitary (see Eq. (5)). Hence, whatever α, Z is a whitened Normal variable, i.e. $h(Z)=\frac{1}{2}\log(2\pi e)$, which is a constant function. In this special case, the inequality in Eq. (14) can be replaced by a strict equality.

6. Impact of the dependence between signals

It has been shown in [4] that there is no spurious minima in $\mathscr{C}_{12}(\mathbf{Z})$ when both S_1 and S_2 are *nearly Gaussian*. Although, to the authors knowledge, there is no similar result neither for $\mathscr{C}_i(Z_i)$, nor for $\mathscr{C}_{12}(\mathbf{Z})$ if the signals are *not* nearly Gaussian.

In this section, the global shape of $\mathscr{C}_1(Z)$ (expressed as a function of the mixing coefficients) where at most one of the two variables is multimodal is analyzed through numerical simulations. The impact of the level of correlation and dependence between S_1 and S_2 on the existence of mixing minima is emphasized.

In the simulations below, the mutual informations and the entropies are computed using the natural logarithm, implying that the corresponding numerical values are expressed in 'nats'. The mixed variables S_1 and S_2 have unit variance. Let us first recall that the $\alpha^2+\beta^2=1$ constraint on the weights guarantees that the mixture has a unit variance ($\sigma_Z^2=1$, see Eq. (5)) only if the original sources S_1 and S_2 are uncorrelated ($\delta_{S_1S_2}=0$, see Eq. (4)); this constraint will be respected in the following.

Mixtures $Z=\alpha S_1+\beta S_2$ of four random variables, illustrated as signals in Fig. 2, are analyzed. These variables have different correlation and dependence levels. The correlation is measured through Pearson's correlation coefficient $\rho(S_1,S_2)$, given for zero-mean variables by

$$\rho(S_1,S_2)=\frac{E\{S_1S_2\}}{\sqrt{E\{S_1S_1^{\mathrm{T}}\}E\{S_2S_2^{\mathrm{T}}\}}} \tag{29}$$

with $0\leqslant|\rho|\leqslant 1$. On the other hand, the mutual information $I(S_1,S_2)$, defined by Eq. (11), measures the statistical independence between the signals ($I(S_1,S_2)\geqslant 0$ with equality if and only if S_1 and S_2 are independent [9]).

Note that while independence implies zero correlation, the reciprocal is false. While a zero correlation means $\delta_{S_1S_2}=0$, independence is reached only if $E\{\varphi(S_1)\psi(S_2)\}=E\{\varphi(S_1)\}E\{\psi(S_2)\}$ for all nonlinear functions φ and ψ [17].

The values of the correlation and mutual information between the signals plotted in Fig. 2 are given in Table 1. These values are *estimated* of the associated theoretical statistical quantities, since the true distributions are supposed to be unknown. For this reason, the estimation of correlation and mutual information depend on the sample size. For instance, only an infinite number of samples allow to get a zero mutual information. The source distributions are estimated with the Parzen estimator and shown in the right column of Fig. 2. Details about this estimator and the choice of the standard deviation of the basis kernels can be found in the appendix.

Fig. 3 shows the entropy of the mixtures as a function of α for several combinations of the signals given in the left column of Fig. 2.

Three results must be pointed out. Firstly, mixing minima do not seem to exist in $\mathscr{C}_1(Z)$ when the original sources S_1 and S_2 are independent (Figs. 3(a) and (b), with sources such that $I(S_1,S_2)$ is small). Secondly, contrarily to the previous case, when the dependence level grows (higher values of $I(S_1,S_2)$), mixing minima appear (Figs. 3(c) and (d)). Note that for correlated variables (with non-negligible ρ, i.e. $\delta_{S_1S_2}\neq 0$), the constraint on the weights given by Eq. (5) no more imposes $\sigma_Z^2=1$ (see Eq. (4)); in this case, a scaled version of Z (defined by $Z^\star=Z/\sigma_Z$) is preferred to ensure the consistency with the source separation problem. If $\delta_{S_1S_2}\neq 0$, a scale factor

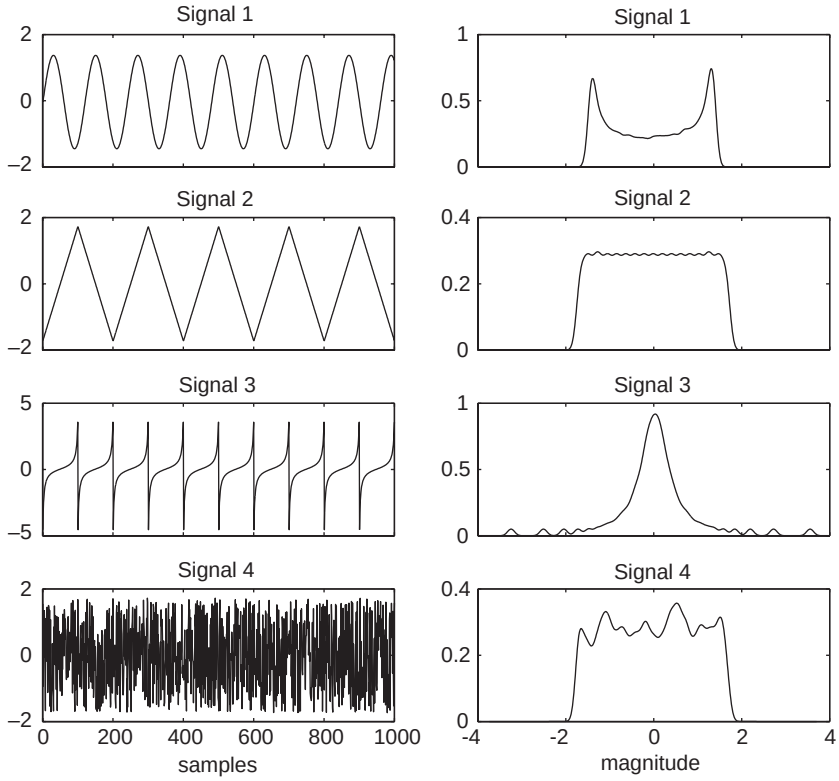

Fig. 2. Signals (left) and their estimated distributions (right).

Table 1
Left: absolute values correlations—$|\rho|$; right: mutual information—I— between signals of Fig. 2 (distributions estimated by Parzen windows)

Signals	1	2	3	4	Signals	1	2	3	4
1	1	**0.08**	**0.06**	0.03	1	1.565	**0.081**	**0.050**	0.020
2	0.08	1	**0.00**	0.00	2	0.081	1.556	**0.385**	0.021
3	0.06	0.00	1	**0.01**	3	0.050	0.385	0.643	**0.019**
4	0.03	0.00	0.01	1	4	0.020	0.021	0.019	1.548

appears between $\mathscr{C}_1(Z)$ and $\mathscr{C}_1(Z^\star)$ (see Eq. (10) and the difference between $\mathscr{C}_1(Z)$—solid—and $\mathscr{C}_1(Z^\star)$—dotted—curves in Fig. 3(c)). On Fig. 3(d) (where independence between sources is not fulfilled), even Eqs. (15) and (16) are violated.

It is obvious from Figs. 3(c) and (d) that we cannot trust in both local and global minima when the sources are mutually dependent; the entropy is not an appropriated cost function for BSS in this case.

Finally, for independent variables with different entropies, $h(Z)$ does not reach its maximum (i.e. more random mixture) for $\alpha = \beta = \frac{\sqrt{2}}{2}$ (see Fig. 3(b)). The most mixed signal, i.e. with equals weights, is thus not necessarily the most random, with the highest entropy.

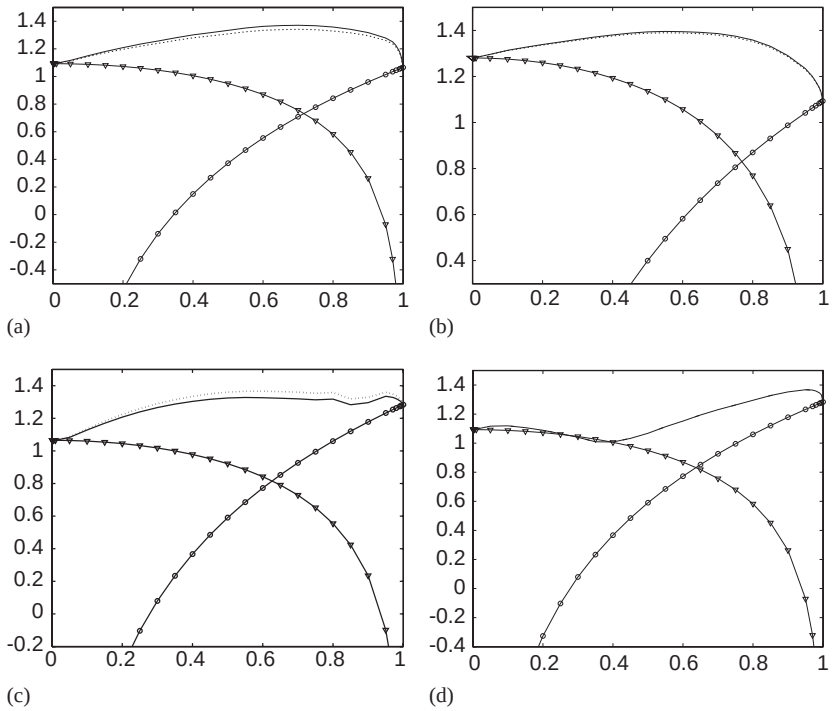

Fig. 3. Entropies vs. α for several whitened variables S_1 and S_2 (see Fig. 2 for labelling). Solid line, $\mathscr{C}_1(Z)$; dashed line, $\mathscr{C}_1(Z^\star)$ where $Z^\star = Z/\sigma_Z$; circles, $h(S_1) + \log|\alpha|$; top-down triangles, $h(S_2) + \log|\beta|$: (a) S_1: signal 1 and S_2: signal 3; (b) S_1: signal 3 and S_2: signal 4; (c) S_1: signal 2 and S_2: signal 1; (d) S_1: signal 2 and S_2: signal 3.

7. Impact of the modality

In the particular case where both the distributions of S_1 and S_2 have more than one mode, the key result observed in the previous case (no mixing minima for $\mathscr{C}_1(Z)$ if $I(S_1, S_2)$ is sufficiently small) vanishes. This is illustrated in the next sections.

7.1. Spurious minima

A first joint density of two bimodal independent variables S_1 and S_2 is shown in Fig. 4(a). In this example, the tails of the modes overlap. The evolution of $\mathscr{C}_1(Z)$ vs. α is given in Fig. 4(b). One can observe a mixing minimum around $\alpha \simeq 0.65$, despite the very low level of dependence between the original sources ($I(S_1, S_2) = 0.016$). It is visible from Figs. 4(c) and (d) that if the modes of the source distributions are not overlapping anymore, this phenomenon can be more clearly emphasized ($I(S_1, S_2) = 0.006$).

By contrast with the previous situation, if each of these sources is combined with any unimodal source from Fig. 2 (signals 2–4), no spurious minimum appears. However, if one of the multimodal sources used in this section is combined with the sine wave from Fig. 2 (bimodal distribution), we can also observe the appearance of a mixing minimum.

Because of these possible mixing local minima, in the particular case of multimodal sources, a gradient descent on $\mathscr{C}_i(Z_i)$ (even if the sources are mutually independent) can lead to spurious minima. In order to emphasize this phenomenon, the graphs in Figs. 4(b) and (d) are plotted vs. θ in Figs. 5(a) and (b), respectively. The non-mixing (acceptable) solutions corresponds to $\theta = k\pi/2\ \forall k \in \mathbb{Z}$ (see Eq. (6)). We can observe that mixing solutions exist around $\theta^\diamond = (2k+1)\pi/4$.

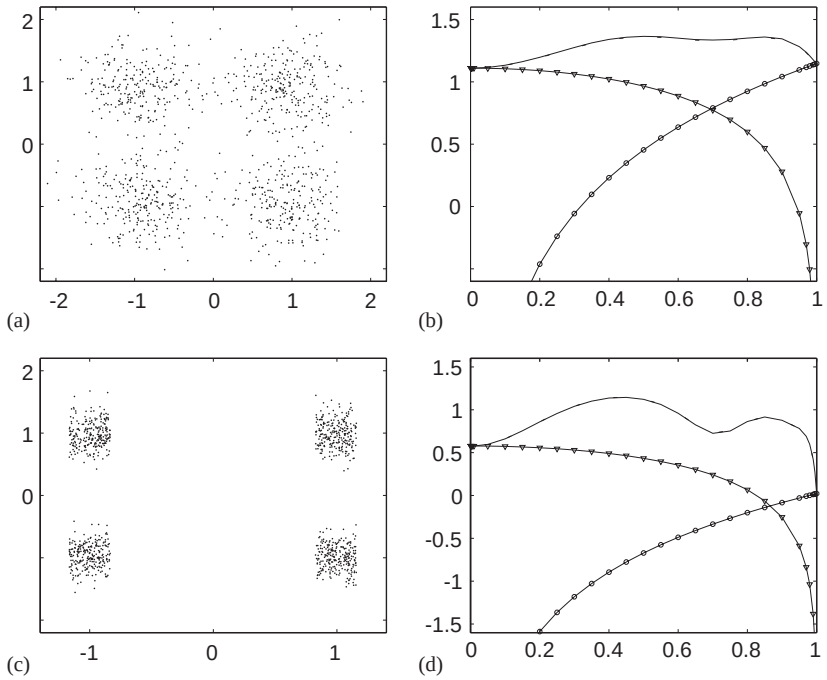

Fig. 4. Effect of the source multimodality on the existence of mixing minima for uncorrelated and quasi-independent sources. Solid line, $\mathscr{C}_1(Z)$; dashed line, $\mathscr{C}_1(Z^\star)$ where $Z^\star = Z/\sigma_Z$; circles, $h(S_1) + \log|\alpha|$; top-down triangles, $h(S_2) + \log|\beta|$: (a) Joint distribution of slightly bimodal independent sources, (b) evolution of $h(Z)$ vs. α, (c) joint distribution of heavy bimodal independent sources, (d) evolution of $h(Z)$ vs. α.

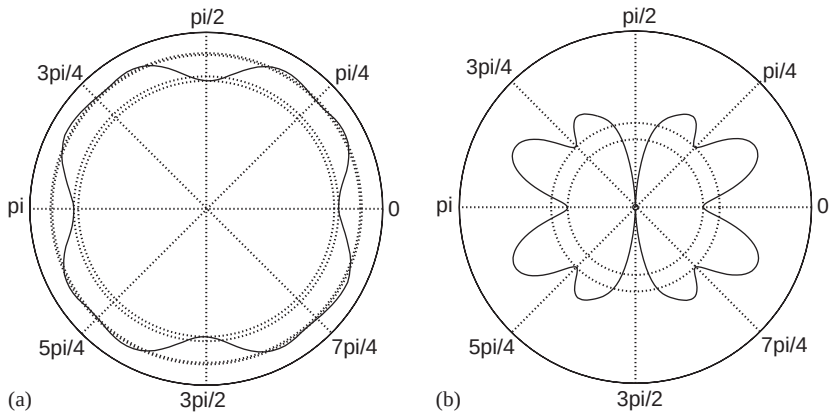

Fig. 5. Evolution of $\mathscr{C}_1(Z)$ vs. θ for uncorrelated and quasi-independent variables having bimodal distributions: (a) Slight bimodality with $h(S_1) \simeq h(S_2)$, (b) high bimodality with $h(S_1) \ll h(S_2)$.

This phenomenon seems to deforce a deflation approach: if $\mathscr{C}_1(Z)$ is trapped in a spurious minimum ($\theta \neq k\pi/2$), the first estimated source will not correspond to an original one; the estimation of the second one will thus fail too, if it is obtained by orthogonalizing the second row of $\mathbf{W}$ with respect to the first one.

Let us denote by w_{ij} the jth entry of the ith row of $\mathbf{W}$. Hence, under the whitening constraint, if $w_{11} \doteq \sin(\theta)$, the system reduces to

$$\mathbf{Z} = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} \sin(\theta) & \cos(\theta) \\ \cos(\theta) & -\sin(\theta) \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \end{bmatrix}. \tag{30}$$

With symmetric algorithms, i.e. where all sources are extracted simultaneously, the cost function is the sum of the marginal entropies [10], as detailed in Section 4.2. One may ask if the local spurious minima also appear with such algorithms. Unfortunately, Fig. 6 shows that even the symmetric approaches do not allow to avoid spurious minima of the associated cost function $\mathscr{C}_{12}(\mathbf{Z})$, and suffer from the same drawback as deflation methods. For this reason, the mutual information criterion may have local minima, when minimized by adjusting an unmixing angle of a rotation matrix. Note that it is shown here that the mutual information may have spurious minima when this criterion is minimized *under constraint that the observations are linear mixtures of the sources signals and that the outputs are a rotated version of the whitened observations*; without this constraint, the mutual information cost function seems to have no local minima (see e.g. [1]).

Remark 1. Note that because of Eq. (16), the mixing minima of $\mathscr{C}_i$ (if they exist) are local, since in this experience both sources respect the assumptions of Corollary 2. Similar conclusions can be drawn regarding $\mathscr{C}_{12}$, since the minimum value (which is zero by Eq. (12)) is reached if and only if the Z_i are mutually independent, i.e. if $\mathbf{Z} = \mathbf{Z}^{\diamond}$ as explained in Section 3.

Hence, both $\mathscr{C}_i$ and $\mathscr{C}_{12}$ can be used as BSS cost functions, since their global minimum corresponds to $\alpha\beta = 0$. Unfortunately, there is no guarantee that their global minima can be reached through a gradient descent when both source variables are multimodal because of the existence of local minima.

Remark 2. The reader can easily check that $h(Z|\theta) = h(Z|\pi + \theta)$, as it can be seen in Fig. 5. This is due to the fact that the sign of Z has no consequence on $\mathscr{C}_1(Z)$ (Eq. (10) with $C = -1$). Moreover, $h(Z|\theta) = h(Z|\pi - \theta)$ if f_{S_1} or f_{S_2} has a symmetry axis. Similarly, the additional symmetries visible in Fig. 6 are due to the fact that permuting Z_1 and Z_2 does not change $\mathscr{C}_{12}(\mathbf{Z})$.

7.2. Discussion

It is interesting to try to explain why such spurious minima in the entropy function appear when multimodal sources are involved in the mixture. Similarly, the values of the critical angles (corresponding to spurious minima of the entropy) should be analyzed: for which reason those minima seem to occur around $\theta^{\diamond}$?

Recall that f_Z is the convolution of $f_{\alpha S_1}$ and $f_{\beta S_2}$ (Eq. (26)). Note that by Eq. (27) the latter

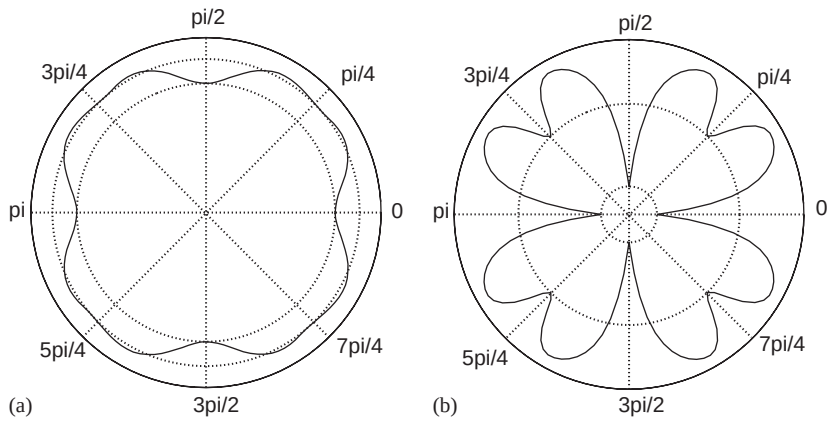

Fig. 6. Evolution of $\mathscr{C}_{12}(\mathbf{Z})$ vs. θ for uncorrelated and quasi-independent variables having bimodal distributions (symmetric approach): (a) Slight bimodality with $h(S_1) \simeq h(S_2)$, (b) high bimodality with $h(S_1) \ll h(S_2)$.

distributions are similar to the densities of the original variables, up to a contraction (resp. extension) of the magnitude (resp. probability) axes.

Consider that f_{S_1} is multimodal with $M>1$ modes and that f_{S_2} is unimodal. The resulting distribution of Z is a distribution having one mode if $|\alpha|$ is small and M modes if $|\alpha|$ is close to one. Therefore when $|\alpha|$ varies from 0 to 1, the structure of f_Z evolves smoothly from an unimodal distribution to a M-modal one.

The case of both multimodal (here bimodal) sources is intrinsically different, the analysis is restricted to angles θ in the first quadrant, but the reasoning holds in the three other ones. An illustration of this case is given in Fig. 7, using the jointly distributed sources from Fig. 4(c). Note that contrarily to σ_Z^2, the variance of αS_1 and βS_2 vary with θ. This implies that a direct estimation of their distributions using fixed-variance basis kernels (see the appendix) is not reliable. In order to illustrate the distribution in Fig. 7, we have circumvent this problem by using Eq. (27) to draw $f_{\alpha S_i}$ as a scaled version of f_{S_i}.

For θ near 0 or $\pi/2$, the structure of f_Z is bimodal (because Z is close from one of the original bimodal variables). This is illustrated in Fig. 7(d) for $\theta=\pi/2$. When θ increases starting from 0, f_Z tends to have a four-modal structure (see Fig. 7(a)), due to the dilatation (resp. contraction) of the X-axis of $f_{\alpha S_1}$ (resp. $f_{\beta S_2}$). The same situation occurs if θ decreases starting from $\pi/2$ (Fig. 7(c)).

Let us denotes $\Delta(S_1,\theta)$ (resp. $\Delta(S_2,\theta)$) the distance between the two modes of $f_{\alpha S_1}$ (resp. $f_{\beta S_2}$) where $\alpha=\sin(\theta)$ and $\beta=\cos(\theta)$. According to this definition, the distances between modes in the original f_{S_1} and f_{S_2} distributions are, respectively $\Delta(S_1,\pi/2)$ and $\Delta(S_2,0)$.

It exists for particular values $\theta^{\square}$ for which the contraction of the axes produces the same distance between modes in $f_{\alpha S_1}|\theta^{\square}$ and $f_{\beta S_2}|\theta^{\square}$: $\Delta(S_1,\theta^{\square})=\Delta(S_2,\theta^{\square})$, as shown in Fig. 7(b). In this case, the convolution of the distributions of αS_1 and βS_2 will be a three-modal density, the

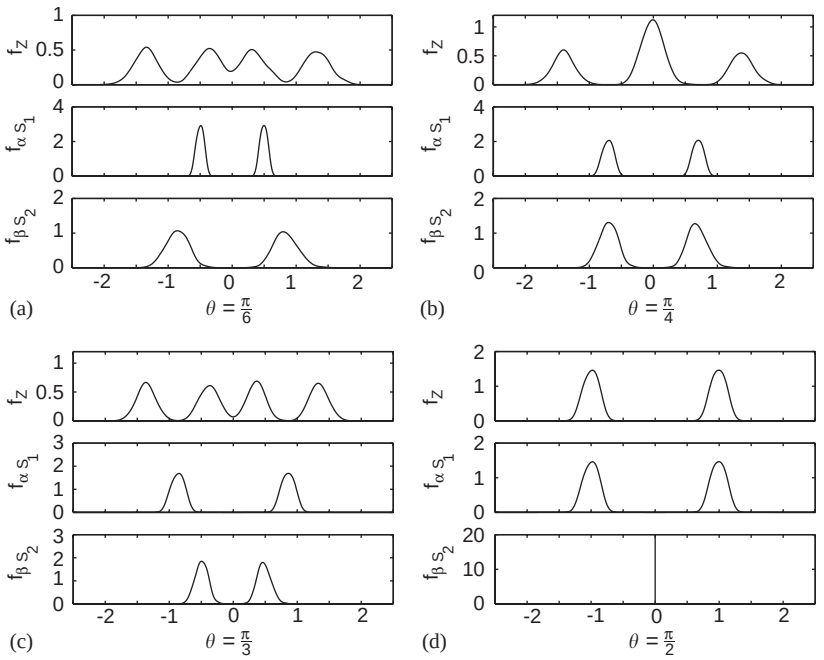

Fig. 7. Effect of θ on the multimodality structure of f_Z and on the contraction of the distributions of $\sin(\theta)S_1$ ($f_{\alpha S_1}$) and $\cos(\theta)S_2$ ($f_{\beta S_2}$): (a) $\pi/6$, (b) $\pi/4$, (c) $\pi/3$, (d) $\pi/2$.

central mode of the result corresponding to the superimposition of $f_{\alpha S_1}|\theta^{\square}$ and $f_{\beta S_2}|\theta^{\square}$; this mode is higher than the two other ones, due to the matching of the two pairs of modes. If the distances between the modes are equal in the original densities ($\Delta(S_1,\pi/2)=\Delta(S_2,0)$), the values $\theta^{\square}$ correspond to the angles verifying $\sin(\theta^{\square})=\cos(\theta^{\square})$.

For the variables used in the example shown in Fig. 7 we have $\Delta(S_1,\pi/2)\simeq\Delta(S_2,0)\simeq 2$ (see their joint distribution plotted in Fig. 4(c)). According to the above development, $\theta^{\square}$ should thus be close to $\pi/4$. This is effectively the case as shown in Fig. 7(b).

Of course, if the distance between the modes of f_{S_1} is different than the one of f_{S_2} ($\Delta(S_1,\pi/2)\neq\Delta(S_2,0)$), then the distances $\Delta(S_1,\theta)$ will become equal to $\Delta(S_2,\theta)$ for a different contraction of their magnitude axes. In other words, a three-modal structure for f_Z is now reached for $\theta^{\square}\neq\pi/4$.

The above justification can intuitively explain the existence of local spurious minima in the entropy function of a linear mixture of multimodal sources. Indeed, keeping a unit-variance mixture and the modal widths approximately constant, if the number of modes in the mixture increases, the entropy (i.e. the 'randomness') of the mixture will increase too. Therefore, if as in the above example there is a local minimum in the number of modes (four then three then four) when α increases, there will also be a local minimum in the entropy function (see Fig. 5(b)).

Consequently, the existence of the spurious minima in the entropy cost function is directly related to a structure modification of f_Z according to the weights. This modification occurs when two multimodal sources are involved in the mixture. The values of the critical angles depend of the distances $\Delta(S_1,\pi/2)$ and $\Delta(S_2,0)$ between the modes of the original distributions.

Nevertheless, it must be stressed that the entropy of a distribution cannot be seen, in all situations, as an increasing function of the number of modes. Indeed, it is not difficult to find examples of whitened variables having a distribution with three small-width modes that have a smaller entropy than other whitened bimodal variables having a distribution with large and close modes. For this reason, the link between the local minima of the entropy and the modality of the distribution is only intuitive. However, it is clear that the mixture entropy extrema are directly related to the particular structure of the mixture distribution, which in turns depend on a specific contraction of the source distributions.

8. Conclusion

Having insights about the global shape of the entropy of a linear mixture of variables is important in order to study the performances of such criterion for source separation. As it is often the case in independent component analysis (ICA), nothing is known about the source distributions but their dependence level. For this reason, theoretical developments of $h(Z)$, where Z is a weighted sum of independent variables S_1 and S_2, will only give few information about the global shape of this function in the general case. Nevertheless, through the entropy power inequality, valid for independent variables, one can prove that the global minimum of $h(Z)$ is reached when only the weight associated to the lowest entropic source is non-zero (under whiteness constraint, its absolute value is equal to one).

According to our numerical simulations, the entropy $h(Z)$ of a mixture of independent variables where at most one has a multimodal distribution seems to have no mixing minimum. This means that in practice, each local minimum of the entropy of a unitary variance weighted sum of independent enough signals S_1 and S_2 is associated to a mixture $Z=Z^{\diamond}$. Hence, it corresponds to one of the original signals: $Z^{\diamond}\propto S_1$ or $Z^{\diamond}\propto S_2$. Therefore a gradient descent on the entropy of a mixture according to its weights will lead to a single non-zero weight: an original source is recovered, up to a real scale factor. If this minimum is global, the lowest entropic source has been extracted.

It is shown that mixing (local) minima may appear even in an ideal mixture of independent

sources. This can occur when the source distributions are multimodal. In [16], Cardoso presents a simple example of spurious minimum when the likelihood-based cost function is used for ICA, even if the distribution of the multimodal sources is a priori known. Similar conclusions are drawn in this paper when the marginal entropy-based cost functions are used instead of the maximum likelihood one. The results also apply when the mutual information criterion is used, since the mutual information and the sum of the marginal entropies are equivalent criteria for ICA under whitening constraint. This result is shown even for joint source distributions having $2 \times 2 = 4$ modes. It is justified by specific contractions of the source distributions (resulting from the weighting) that influence the number of modes in the mixture. The justification also allows us to justify the locations of the mixing minima knowing some information about the modality of the source distributions.

In the case of mutually dependent signals, it is shown that local mixing minima may appear whatever is the number of modes in the source distributions. Moreover, the global minimum of the mixture entropy does not necessarily correspond to a non-mixing solution, since there is no guarantee that the entropy power inequality holds. This explain why the entropy function cannot be used as cost function in this case.

Finally, it is observed that the maximum entropy for mixtures of independent variables is not necessarily reached for the most mixed signal (equal weights in the mixture), but depends of the entropy of the original sources and their number of modes.

To conclude, even if it is true that a Gaussian variable has the maximum entropy among all the unbounded variables of same variance, the naive "*mixing gaussianizes*" interpretation of the central limit theorem must be taken with care.

Acknowledgements

The authors would like to thank Jean-François Cardoso for valuable discussions and for inspiring investigations on multimodal sources.

Appendix. Estimation of distributions by the Parzen method

In this appendix, the Parzen estimator [18] for one-dimensional distributions is recalled. The extension to higher dimensions is trivial. In particular, it is shown how the standard deviation of the basis Gaussian kernel has been chosen in the experiments.

The *Parzen estimator* is one of the most known (and simple) methods to estimate easily the distribution of a variable X from a finite number N of samples. Consider a simple isotropic kernel $\mathscr{K}$ with mean μ and standard deviation σ such that the $\int \mathscr{K}(x, \mu, \sigma)\,\mathrm{d}x = \eta$. The density estimation consists in placing a basis kernel $\mathscr{K}$ of fixed variance σ^2 on each point $X(i)$ in the magnitude space of X. The estimation $p(x)$ of the distribution is the sum of these basis functions, up to a scale factor, in order to guarantee $\int p(x)\,\mathrm{d}x = 1$:

$$p(x) \doteq \frac{1}{N\eta} \sum_{i=1}^{N} \mathscr{K}(x, X(i), \sigma). \tag{31}$$

Particular attention must be paid to the choice of the standard deviation of the basis kernel, in order to avoid under- and over-fitting. In this paper, isotropic Gaussian kernels are used.

To choose an adequate value for the standard deviation parameter, a simple heuristic was developed. It is well known that the third signal (noted here $S3$) in Fig. 2 (triangular temporal structure) has a uniform distribution. Consequently, the optimal value of the σ parameter was taken as the value $\sigma = \sigma^\star$ that minimizes the $\mathscr{L}_1$ norm between the estimated distribution $p(S3|\sigma)$ of the whitened triangular signal and the unit-variance uniform distribution (noted $\mathscr{U}$):

$$\sigma^\star = \arg\min_{\sigma} |p(S3|\sigma) - \mathscr{U}|. \tag{32}$$

This optimal value is close to 0.08 (see Fig. 8). This parameter influences all numerical data involving the estimated distributions (e.g. the entropy and the mutual information). For this reason, the same standard deviation $\sigma = \sigma^\star$ has been used in the basis kernels to estimate all distributions in this paper, in order to avoid any undesired effect of such choice.

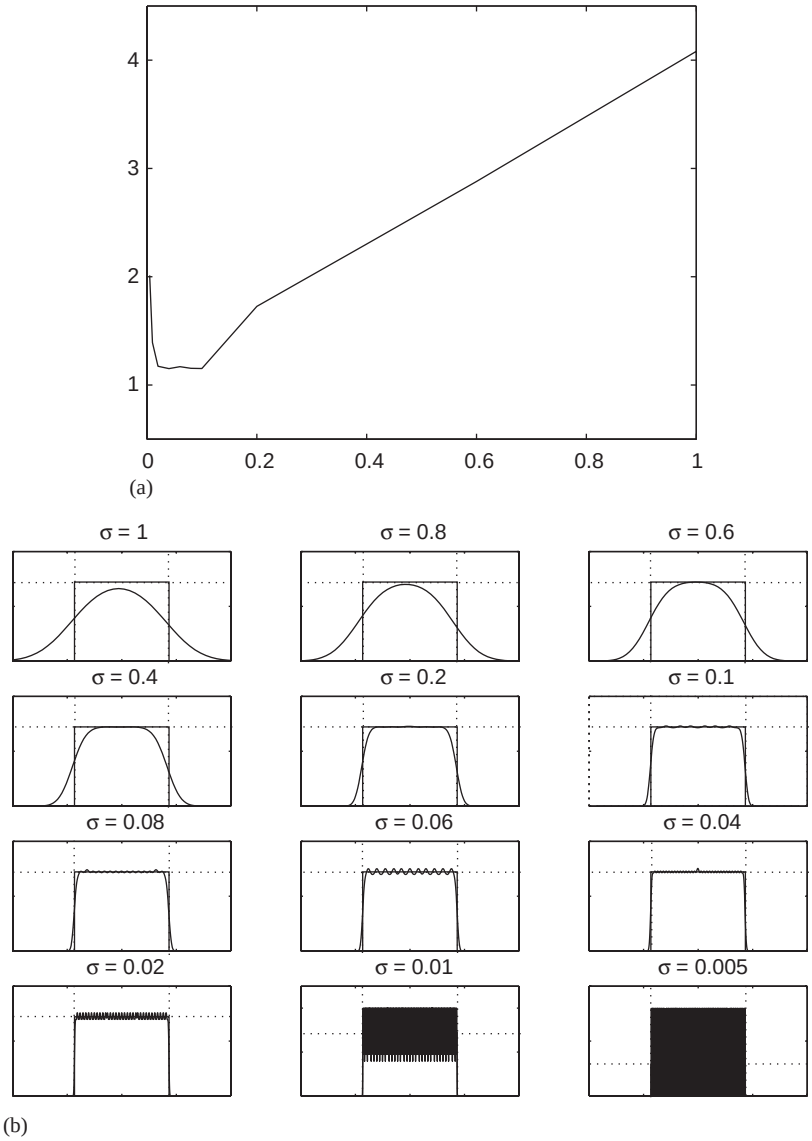

Fig. 8. Impact of the variance of the isotropic basis Gaussian kernels for uniform distribution estimation: (a) Evolution of the $\mathscr{L}_1$ norm between the estimated distribution p and the true uniform distribution U, (b) estimated distribution p and unit-variance uniform distribution U for various values of σ.

References

[1] M. Babaie-Zadeh, C. Jutten, A general approach for mutual information minimization and its application to blind source separation, Signal Processing, this issue.

[2] A.J. Bell, T.J. Sejnowski, An information-maximisation approach to blind separation and blind deconvolution, Neural Comput. 7 (6) (1995) 1129–1159.

[3] R. Boscolo, H. Pan, V. Roychowdhury, Independent component analysis based on nonparametric density estimation, IEEE Trans. Neural Networks 15 (1) (2004) 55–65.

[4] R. Boscolo, V. Roychowdhury, On the uniqueness of the minimum of the information-theoretic cost function for the separation of mixtures of nearly Gaussian signals, in: Proceedings of the International Conference on

Independent Component Analysis and Blind Signal Separation (ICA'03), 2003, pp. 137–141.
[5] J.-F. Cardoso, Source separation using higher order moments, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'89). Glasgow (England), 1989.
[6] J.-F. Cardoso, Blind signal separation: statistical principles, special issue on blind identification and estimation, in: R.-W. Liu, L. Tong (Eds.), Proceedings of the IEEE, 1998, pp. 2009–2025.
[7] J.-F. Cardoso, Dependence, correlation and gaussianity in independent component analysis, J. Mach. Learning Res. 4 (2003) 1177–1203.
[8] P. Comon, Independent component analysis, a new concept?, Signal Processing 36 (3) (1994) 287–314.
[9] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, New York, 1991.
[10] S. Cruces, A. Cichocki, S. Amari, The minimum entropy and cumulants based contrast functions for blind source extraction, in: J. Mira, A. Prieto (Eds.), Proceedings of the International Workshop on Artificial Neural Networks (IWANN'01), Lecture Notes in Computer Science, vol. 2085, Springer, Berlin, 2001, pp. 786–793.
[11] S. Cruces, A. Cichocki, S. Amari, From blind signal extraction to blind instantaneous signal separation: criteria, algorithms and stability, IEEE Trans. Neural Networks 15 (4) (2004) 859–873.
[12] G. Darmois, Analyse générale des liaisons stochastiques, Rev. Inst. Int. Stat. 21 (1953) 2–18.
[13] N. Delfosse, P. Loubaton, Adaptibe blind separation of sources: a deflation approach, Signal Processing 45 (1995) 59–83.
[14] D. Donoho, On minimum entropy deconvolution, in: D. Findley (Ed.), Applied Time Series Analysis II, Academic Press, New York, 1981, pp. 565–608.
[15] B. Godfery, An information theory approach to deconvolution, Technical Report 15, Stanford Exploration project, 1978.
[16] S. Haykin (Ed.), Unsupervised Adaptive Filtering, vol. 1, Blind Source Separation, Wiley, New York, 2000.
[17] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York, 2001.
[18] E. Parzen, On estimation of a probability density function and mode, Ann. Math. Stat. 33 (1962) 1065–1076.
[19] D.-T. Pham, Blind separation of instantaneous mixtures of sources based on order statistics, IEEE Trans. Signal Process. 48 (2) (2000) 363–375.
[20] J. Rice (Ed.), Mathematical Statistics and Data Analysis, Duxurby Press, Belmont, CA, 1995.
[21] C.E. Shannon, The Mathematical Theory of Communication The Board of Trustees, University of Illinois, Urbana, IL, 1949.