Entropy Minima and Distribution Structural Modifications in Blind Separation of Multimodal Sources

Frédéric Vrins, Cédric Archambeau and Michel Verleysen[1]

Université catholique de Louvain (UCL) - Machine Learning Group
Place du Levant 3, 1348 Louvain-la-Neuve, Belgium
{vrins, archambeau, verleysen}@dice.ucl.ac.be

Abstract.
The source separation problem is usually solved through a gradient descent on a cost function $\mathscr{C}$. However, $\mathscr{C}$ may have local minima that are irrelevant from the source separation point of view in particular when the source distribution is multimodal. Cardoso explained the reason for such spurious minima when a likelihood-based function is used as cost criterion, even when the source distributions are a priori known.

This paper shows that such spurious minima may also appear when using the *marginal entropy* cost function; it aims to draw an intuitive justification about the existence and the locations of these minima when dealing with multimodal sources. This justification is based on a structural modification (mainly the modality) analysis of the output distribution according to the mixing coefficients.

INTRODUCTION

In order to solve the blind source separation (BSS) problem, independent component analysis (ICA) can be used if some assumptions are met. ICA tries to find linear combinations of measured signals in order to produce *output signals* as independent as possible. Several cost functions $\mathscr{C}$ were derived to measure the dependence level between outputs. This paper focusses on one of them: the sum of the marginal entropies of the output signals [1]. The solution to the ICA problem (the optimal linear combinations of the measured signals) is found through the minimization of a criterion measuring the dependence between the output signals. The minimum is usually reached through a gradient descent on $\mathscr{C}$. However, this gradient descent process is meaningful if and only if all local minima of $\mathscr{C}$ are relevant from the source separation viewpoint.

A well-known way to perform independent component analysis is to use as cost function $\mathscr{C}$ the Kullback-Leibler (KL) divergence between an assumed model for the original source distribution (called *target distribution*) and the output distribution. In [2], Cardoso shows that spurious minima in this measure appear when the marginal distributions of the sources are multimodal. Recently, it was noted by several authors that the entropy

[1] MV is Senior Research Associate with the Belgian National Funds for the Scientific Research (FNRS). This work was partially supported by the European Commission (IST-2000-25145).

CP735, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*,
Edited by R. Fisher, R. Preuss, and U. von Toussaint

cost function may also have spurious minima in this context [3, 4, 5]. However, since the entropic approach does not suppose any model for the source distribution, the existence of spurious minima cannot be understood by the same arguments as in the KL case.

This paper aims to explain how spurious minima may appear when the (sum of the) output marginal entropy(ies) is used as cost function on a multimodal BSS problem. This is done by looking to the effects of scaling and mixing independent random variables. Furthermore, this analysis allows to understand the locations of the possible spurious minima, i.e. for which mixture coefficients they appear.

The remaining of the paper is organized as follows. First, the mixing process of two variables is recalled. Next, the blind source separation (BSS) problem and the independent component analysis (ICA) method are detailed. Using a simple example, we illustrate that spurious minima in the marginal entropy cost function may appear if the source distributions are multimodal. Finally, we explain *why* and *where* spurious minima appear in the entropy function, leading to bad solutions in the source separation problem when using a gradient descent algorithm.

LINEAR MIXTURE OF RANDOM VARIABLES

In this section, we focus on the mixing process of independent, stationary, ergodic and real variables, and study its effect on the distribution of the resulting variables.

The whitened mixture scheme

In many real-world applications involving signals, the sensor recordings correspond to mixtures of original sources. For example, consider that $m \geq 2$ speakers speak in a room and that m microphones, located at different places, record the ambient noise. Each microphone does not only record the speech of a single speaker, but the whole acoustic signal emitted simultaneously by all the m acoustic sources (i.e. the m speakers). The problem of separating the acoustic sources from the sensor signals is known in this case as the 'cocktail party' problem. Before looking at the source separation process, it is necessary to analyze the mixture scheme of independent random variables. We assume that the m recorded signals $\mathbf{X}(t) = [X_1(t), \ldots, X_m(t)]^T$ resulting from a mixing process can be modelled by a linear combination of m sources $\mathbf{S}(t) = [S_1(t), \ldots, S_m(t)]^T$:

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

where T denotes the transposition and $\mathbf{A}$ a real $m \times m$ mixing matrix, constant in time. In the following, we will omit the temporal variable t, for the simplicity of presentation.

A useful preprocessing to ICA is to center and whiten the sensor signals $\mathbf{X}$ such that they are zero-mean and have an identity covariance matrix: $E\{\mathbf{X}\} = 0$ and $E\{\mathbf{X}\mathbf{X}^T\} = \mathbf{I}_m$, where E denotes the statistical expectation and $\mathbf{I}_m$ the identity matrix of size $m \times m$. Without loss of generality, we also assume that the sources are zero-mean and have unit variance. It can be shown that these constraints imply that $\sum_j a_{ij}^2 = 1$, where a_{ij} denote the elements of $\mathbf{A}$.

Weighting and summing variables

It is well known that the distribution f_V of a variable $V = \alpha U$ ($\alpha \in \mathbb{R}$) is directly linked to the distribution f_U of U by the following relation:

$$f_{\alpha U}(v) = \frac{1}{|\alpha|} f_U\left(\frac{v}{\alpha}\right) \ . \tag{2}$$

Since $\int f_U(u)du = 1$, if the maximum value of f_U increases (resp. decreases), the support Ω_u of f_U is contracted (resp. extended). Of course, this seems to be a non-sense if Ω_u is infinite. Actually, this contraction/extension of the support should be understood considering 'the inter-distances' between the elements of Ω_u (see below). Multiplying a variable by a real scaling coefficient smaller (resp. greater) than one contracts (resp. extends) the support of the distribution.

Another interesting fact is that the distribution f_Z of $Z = U + V$ where U, V are two independent variables is the convolution of f_U and f_V [6]:

$$f_Z(z) = f_U * f_V = \int f_U(\tau) f_V(z-\tau) d\tau \ . \tag{3}$$

BLIND SOURCE SEPARATION

The problem of source separation consists in recovering the original signals $\mathbf{S}$ knowing only $\mathbf{X}$; the sources are recovered applying an unmixing matrix $\mathbf{B}$ on $\mathbf{X}$. Under the assumptions of the Darmois-Skitovich theorem (DST) [1], this can be done using independent component analysis (ICA). In 1994 [7], Comon has shown that we can only recover $\mathbf{Y} = \mathbf{BX} = \mathbf{BAS} = \mathbf{PDS}$, where $\mathbf{P}$ and $\mathbf{D}$ are permutation and scaling matrices, respectively. The global transfer matrix $\mathbf{W}$, defined by $\mathbf{Y} = \mathbf{WS}$, is thus equal to $\mathbf{BA}$. In the following, the output signals $\mathbf{Y}$ are also supposed to have an identity covariance matrix.

Under the DST assumptions, finding pairwise independent output signals is equivalent to recover signals that are proportional to the original sources. In order to find such independent outputs, ICA algorithms necessitate a measure of dependence. The latter can be the 'distance' between the output (joint) distribution $f_{\mathbf{Y}}$ and the product of the marginal ones $\prod_{i=1}^{m} f_{Y_i}$. Using simple algebraic relations, it can be shown that under the whitening constraint of $\mathbf{Y}$, minimizing the Kullback-Leibler divergence (KL) between $f_{\mathbf{Y}}$ and $\prod_{i=1}^{m} f_{Y_i}$ (also known as the mutual information between the f_{Y_i} [8]) is equivalent to minimizing the sum of the output marginal entropies $H(Y_i)$:

$$\mathscr{C}(\mathbf{Y}) = KL\Big(f_{\mathbf{Y}} || \prod_{i=1}^{m} f_{Y_i}\Big) = \sum_{i=1}^{m} H(Y_i) \ , \tag{4}$$

where $H(Y_i)$ denotes the Shannon entropy of variable Y_i. This last quantity is defined as:

$$H(Y_i) = -\int f_{Y_i} \log(f_{Y_i}) \ , \tag{5}$$

(with $0\log(0) \doteq 0$). The global minimum of $\mathscr{C}(\mathbf{Y})$ is known to be an acceptable solution from the BSS point of view [9]. To reach the minimum value of $\mathscr{C}$, most ICA algorithms use a gradient descent on $\mathscr{C}(\mathbf{Y})$ in order to avoid an exhaustive search. Doing so, one implicitly supposes that all (local) minima are also meaningful results from the source separation point of view. Unfortunately, this is not the case in several situations, as it will be shown in the following.

DEALING WITH MULTIMODAL SOURCES

Dealing with multimodal sources in BSS is known to be a difficult problem, when achieved through a gradient descent on a cost function. Indeed, the usual cost functions used in the ICA algorithms may have spurious minima in such situations; the only alternative to gradient descent is the exhaustive search [4]. The following section recalls the existence of spurious minima when the 'maximum-likelihood' approach is used. Next, similar conclusions will be drawn regarding the entropic cost function.

Spurious minima in the negative likelihood cost function

The maximum-likelihood (ML) approach to BSS consists in finding an output distribution that is as close as possible to a target distribution, which is supposed to be – very close from – the unknown source distribution. The ML-based function has local minima if the marginal source distributions are multimodal [2], even if the target distribution is taken exactly equal to the (unknown) source distribution. These local minima are due to a local optimal matching (in the KL divergence sense) between the output and the target distributions. However, this justification cannot be extended to the marginal entropy criterion, since in this case there are no target distributions.

Spurious minima in the marginal entropy cost function

Assume that the sources S_i are ordered with respect to their Shannon's entropy: $H(S_1) \leq \ldots \leq H(S_m)$. The global minima of $H(Y_k)$ $(1 \leq k \leq m)$ are known to be an acceptable solutions: the output Y_k corresponds to S_1, the lowest entropic source [1]. Local minima appear when Y_k corresponds to S_j with $j \neq 1$, but spurious minima, that are not relevant for BSS, may also appear. This is e.g. the case when dealing with independent sources, linearly and simultaneously combined, without additive noise, but having multimodal distributions. In this section, we will show a simple example of spurious minima in $\mathscr{C}(\mathbf{Y})$.

Consider two independent sources S_1 and S_2 with multimodal distributions (see Fig. 1) and the following global transfer system, with w_{ij} the transfer coefficients:

$$\begin{cases} Y_1 &= w_{11}S_1 + w_{12}S_2 \\ Y_2 &= w_{21}S_1 + w_{22}S_2 \end{cases} . \tag{6}$$

In the remaining of the paper, we will focus on the first entry of $\mathbf{Y}$. It can be shown that the $w_{i1}^2 + w_{i2}^2 = 1$ condition ensures that Y_i is white if $\mathbf{X}$ is white [10]. For this reason, under the whitening constraint, Y_1 can be rewritten without loss of generality as follows:

$$Y_1 = \underbrace{\sin(\theta)S_1}_{S_1^\theta} + \underbrace{\cos(\theta)S_2}_{S_2^\theta} \ , \tag{7}$$

where $\sin(\theta) \doteq w_{11}$ and S_1^θ and S_2^θ are defined in the equation. According to equation 3, the distribution of Y_1 is the convolution of the distributions of S_1^θ and S_2^θ:

$$f_{Y_1} = f_{S_1^\theta} * f_{S_2^\theta} \ . \tag{8}$$

We will adopt the following notation. The distance between two modes i and j of f_{S_k} will be noted $\Delta_{i,j}(S_k)$. For example, we can see in Fig. 1 that $\Delta_{1,2}(S_1) \simeq 2, \Delta_{1,2}(S_2) \simeq 1.1$ and $\Delta_{2,3}(S_2) \simeq 1.6$.

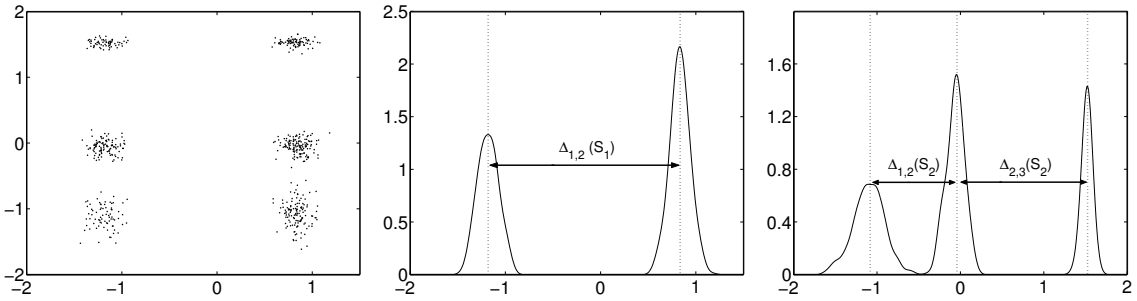

FIGURE 1. Characteristics of the source signals S_1 and S_2: scatter plot f_{S_2} vs f_{S_1}(left), f_{S_1} (center) and f_{S_2}(right).

The entropy minima analysis is restricted to $\theta \in [0, \pi/2]$; the extension to the other quadrants is trivial. The solid curve on the left graph of Fig. 2 shows the evolution of Y_1 vs θ. The only minima relevant for source separation (see eq. 7) correspond to $w_{11}w_{12} = 0$, i.e. to $\theta \in \{0, \pi/2\}$. As it can be seen on Fig. 2, spurious minimum appear for $\theta \neq \{0, \pi/2\}$; this is the case for several angles $\theta^{\maltese} \simeq \{\pi/6, \pi/5, 11\pi/36\}$. These minima are thorny because in these cases, Y_1 remains a mixture of the sources ($w_{11}w_{12} \neq 0$); they correspond to spurious solutions. As previously explained, these spurious minima are local; the global minimum of $H(Y_1)$ is reached when $Y_1 = S_1$, i.e. for $\theta^\star = \arg\min_\theta H(Y_1) = \pi/2$ [9].

The distributions are estimated nonparametrically using a Parzen Window estimator [11] with Gaussian isotropic kernels of standard deviation $\sigma_K = 0.05$. The standard deviation σ_K of the kernels may influence the quality of the estimated distribution. However, it seems that this is not the case (in a certain range) regarding the shape of the entropy function; the latter is shown on the left panel of Fig. 2 where $f_{Y_1}(y_1)$ is plotted vs θ for σ_K=0.025, 0.05 and 0.1. In order to improve the readability of this function, it has been plotted on a polar graph (right panel of Fig. 2). As the radius denotes the entropy, negative entropies cannot be shown; for this reason, $H(Y_1)$ has been shifted to $H(Y_1) + \varepsilon$, where

$$\varepsilon = \begin{cases} 0 \text{ if } \min_\theta H(Y_1) \geq 0 \\ -\min_\theta H(Y_1) \text{ if } \min_\theta H(Y_1) < 0 \end{cases} . \tag{9}$$

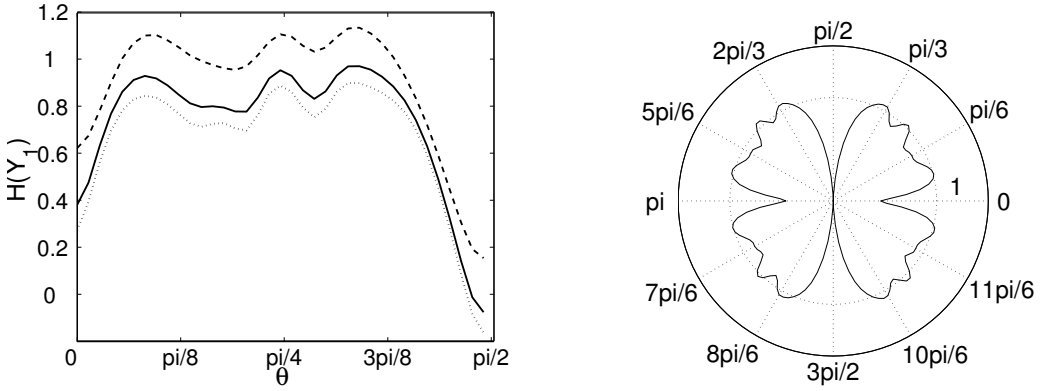

FIGURE 2. Left: Entropy $H(Y_1)$ vs θ for $\sigma_K = 0.025$ (dotted), 0.05 (solid) and 0.1 (dashed); Right: $H(Y_1)+\varepsilon$ vs θ ($\sigma_K = 0.05$).

In the next section, we justify why this phenomenon appears in the particular case of multimodal source distributions. This justification allows also to understand, knowing the source distributions, where these minima are located.

EFFECT OF MIXING VARIABLES ON THE RESULTING DISTRIBUTION

As previously explained, minima in $H(Y_1)$ for $\theta \in]0,\pi/2[$ are spurious. In order to understand why such spurious minima appear, it is useful to plot the evolution of f_{Y_1}, $f_{S_1^\theta}$ and $f_{S_2^\theta}$ vs θ. This is done in Fig. 3 for $\theta = \{0, \pi/12, \pi/6, \pi/5, \pi/4, 11\pi/36, 13\pi/36, \pi/2\}$.

We can observe that the critical values θ^Φ of θ, corresponding to the spurious minima of $H(Y_1)$ also minimize locally the number $N(Y_1)$ of modes of f_{Y_1}. This fluctuation of $N(Y_1)$ as a function of the angle θ (i.e. as a function of the transfer coefficients w_{1i}) is due to the joint effect of the scaling and the mixing of the independent sources S_1 and S_2. Obviously, looking to equation 7, $N(Y_1)$ is equal to $N(S_1)$ (resp. $N(S_2)$) if $\theta = \pi/2$ (resp. 0). Mixing these independent sources (keeping the variance of the mixtures unitary) has for effect to convolute the scaled densities. Intuitively, as $N(S_1) = 2$ and $N(S_2) = 3$, when θ increases from 0 or decreases from $\pi/2$, $N(Y_1)$ should be equal to 6. However, $N(Y_1)$ is not strictly increasing to a unique local maximum when θ moves apart from $k\pi/2$. The function $N(Y_1)$ has several local maxima for $\theta \in [0,\pi/2]$ and $\mathscr{N} \equiv \{2,3,6\}$ is not the whole set of acceptable values for $N(Y_1)$; f_{Y_1} may have (locally) a particular structure if the intermodal distances of distributions $f_{S_1^\theta}$ and $f_{S_2^\theta}$ become equal. In this case, $N(Y_1) < 6$ since two pairs of modes are superimposed during the convolution process. This situation occurs for several scaling factors of S_1 and S_2.

As an illustration, consider the case of $\theta = \pi/5$. This particular angle has the remarkable property to contract the distributions f_{S_1} and f_{S_2} such that $\Delta_{1,2}(S_1^{\pi/5}) \simeq \Delta_{2,3}(S_2^{\pi/5})$.

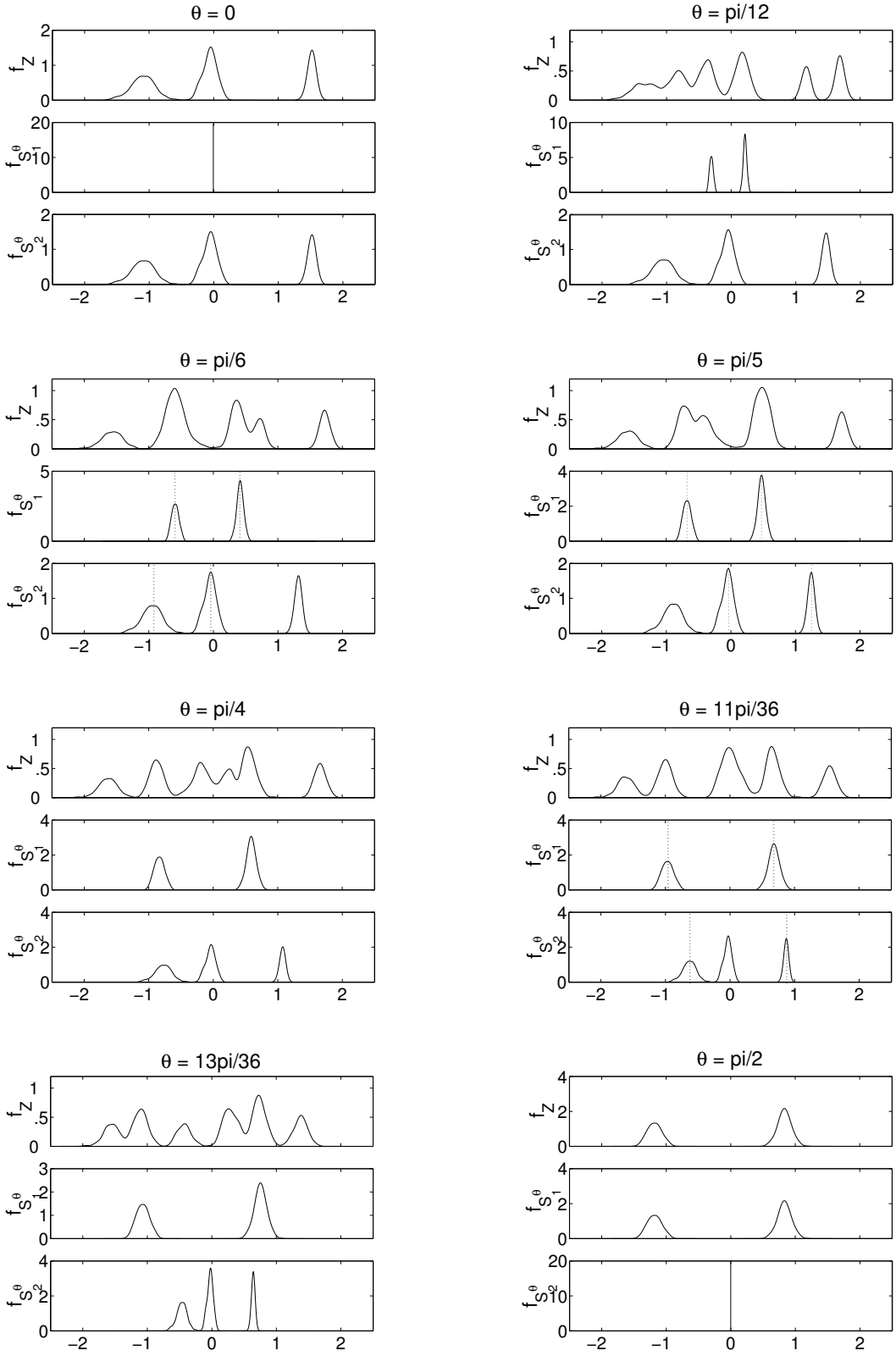

FIGURE 3. Distributions f_{Y_1}, $f_{S_1^\theta}$ and $f_{S_2^\theta}$ for several values of θ.

The distribution of $Y_1 = S_1^{\pi/5} + S_2^{\pi/5}$ results from the convolution of $f_{S_1^{\pi/5}}$ and $f_{S_2^{\pi/5}}$. Due to the matching of the two modes of $f_{S_1^{\pi/5}}$ and the two last modes of $f_{S_2^{\pi/5}}$, the number of modes of Y_1 decreases: $N(Y_1) = 5$. The same phenomenon appears for other values of θ^{Ψ} : $\Delta_{1,2}(S_1^{\pi/6}) \simeq \Delta_{1,2}(S_2^{\pi/6})$, $\Delta_{1,2}(S_1^{11\pi/6}) \simeq \Delta_{1,3}(S_2^{11\pi/6})$. This structural modification of f_{Y_1} (appearing locally around θ if $\theta \in \theta^{\Psi}$) implies a variation of the entropy.

Note that in general, the relation that links the entropy of a variable to the number of modes of its distribution is not so simple: counter examples may be found easily, adjusting the width of the modes. Nevertheless, it is emphasized in this paper that when comparing normalized distributions f_Y *resulting from the convolution of two scaled versions of given distributions* f_{S_1} and f_{S_2}, the modality of f_Y is related to the entropy of Y: they vary similarly when modifying the mixture weights.

CONCLUSION

This paper focuses on the marginal entropy of a sphered linear mixture of two independent source signals. The existence of spurious minima in this entropic function when the sources have multimodal distributions is emphasized. The local minima are due to the structural modifications of the mixture distribution f_Y. The paper shows that the number of modes of f_Y is a function of i) the modality of the source signals and ii) the transfer coefficients. Knowing the source distributions, it is possible to predict for which values of the transfer coefficients local minima in the number of modes will appear. These local minima correspond to the local minima of the entropic cost function, which are a consequence of the convolution of the original source distributions. Therefore, using any gradient-based method on entropy-based independence criteria may lead to false solutions to the Independent Component Analysis problem. Future work should address this problem for $n > 2$.

REFERENCES

1. Cruces, S., Cichocki, A., and Amari, S., "The minimum entropy and cumulants based contrast functions for blind source extraction," in *IWANN'01, LNCS 2085*, edited by J. Mira and A. Prieto, Springer-Verlag, 2001, pp. 786–793.
2. Haykin, S., editor, *Unsupervised Adaptive Filtering vol.1 : Blind Source Separation (ch. IV, pp 171-173)*, John Willey ans Sons, Inc., New York, 2000.
3. Boscolo, R., Pan, H., and Roychowdhury, V., *IEEE Trans. on Neural Networks*, **15**, 55–65 (2004).
4. Learned-Miller, E. G., and Fisher III, J. W., *Journal of Machine Learning Research*, **4**, 1271–1295 (2003).
5. Vrins, F., and Verleysen, M., *submitted for publication to Signal Processing*.
6. Hirschman, I., and Widder, D., *The convolution transform*, Princeton University Press, 1955.
7. Comon, P., *Signal Processing*, **36**, 287–314 (1994).
8. Cover, T. M., and Thomas, J. A., *Elements of information theory*, Wiley and sons, 1991.
9. Cruces, S., Cichocki, A., and Amari, S., *IEEE Trans. on Neural Networks*, **15**, 859–873 (2004).
10. Hyvärinen, A., Karhunen, J., and Oja, E., *Independent component analysis*, John Willey ans Sons, Inc., New York, 2001.
11. Parzen, E., *Ann. Math. Stat*, **33**, 1065–1076 (1962).