EUSIPCO 05 - 13th European Signal Processing Conference
Antalya (Turkey), 4-8 September 2005, pp. 1107:1-4

CAN WE ALWAYS TRUST ENTROPY MINIMA IN THE ICA CONTEXT ?

Frédéric Vrins, John A. Lee and Michel Verleysen

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

ABSTRACT

Marginal entropy can be used as cost function for blind source separation (BSS). Recently, some authors have experimentally shown that such information-theoretic cost function may have spurious minima in specific situations. Hence, one could face spurious solutions of the BSS problem even if the mixture model is known, exactly as when using the maximum-likelihood criterion. Intuitive justifications of the spurious minima have been proposed, when the sources have multimodal densities.

This paper aims to give mathematical arguments, complementary to existing simulation results, to explain the existence of such minima. This is done by first deriving a specific entropy estimator. Then, this estimator, although reliable only for multimodal sources with small-overlapping Gaussian modes, allows one to show that spurious minima may exist when dealing with such sources.

1. INTRODUCTION

In [1, 2], the output marginal entropy (also called Shannon's entropy) is used as cost function to develop a deflation approach to blind source separation (BSS). Shannon's entropy of a random variable U is defined by [3]

$$H(U) = -\int_{\Omega_U} p_U(\xi)\log p_U(\xi)d\xi \ , \qquad (1)$$

where Ω_U and p_U denote the support and the probability density function (pdf) of U, respectively. In this paper, the basis of the logarithm is taken equal to 2.

Under the common independent component analysis (ICA) assumptions, globally minimizing the entropy of a unit-variance instantaneous linear mixture y_i of n independent sources $s_1, \ldots, s_n$ leads to recovering the lowest entropic source [2]. The entropy of an output y_i can be rewritten as

$$H(y_i) = H\left(\sum_j \mathbf{c}_i(j)s_j\right) \text{ s.t. } 1 \leq i,j, \leq n \ , \qquad (2)$$

where $\mathbf{c}_i$ denotes the i-th row of the transfer matrix.

The marginal entropy is an interesting alternative to mutual information, since according to (1), it only requires a one-dimensional density estimation. When using (2) as ICA cost function, a pre-whitening step may be used. Hence, if we constrain the output signal to be whitened, the unmixing matrix must be orthogonal: it must lay on the Stiefel manifold [4]. This manifold is the $n(n-1)/2$ dimensional subspace spanned by the orthogonal matrices in the whole $n\times n$ dimensional space of the square matrices.

One must distinguish two kinds of entropy minima. The *non-mixing* ones refer to $H(y_i)$ local minima in which $|\mathbf{c}_i|$ (the vector composed of the absolute value of $\mathbf{c}_i$'s entries) is equal to a row of $\mathbf{I}_n$ (the $n \times n$ identity matrix). On the contrary, the *mixing* ones correspond to entropy minima for which at least two entries of $\mathbf{c}_i$ are non-zero. In the source extraction application, all non-mixing minima of $H(y_i)$ are satisfactory solutions. By contrast, all mixing minima, if they exist, are spurious. Recently, several results have shown that in some cases, $H(y_i)$ has mixing minima (see [1] and references therein). The existence of such minima, appearing when dealing with multimodal sources, has been understood by looking to the modality of the output distribution, which is a function of the mixture weights. Nevertheless, the link between modality and entropy was only justified by intuitive considerations, and not by mathematical arguments.

The aim of this paper is twofold. First, we derive an entropy estimator for variable having multimodal distribution with small overlap between the modes. Second, this estimator is applied to present mathematical arguments showing that spurious minima in the marginal entropy cost function may exist. The paper is organized as follows. After some definitions, the entropy approximator is derived in Section 2: a simple example shows its efficiency to estimate the entropy of a variable having a distribution with several non-overlapping Gaussian modes (NOGM). For such variables, the estimator is used to link modality and entropy. Next, in Section 4, we focus on the $n = 2$ BSS application: two sources with distributions having two NOGM are presented. The entropy estimator is then applied to compare $H(\mathbf{c}_i(1)s_1 + \mathbf{c}_i(2)s_2)$ for various $\mathbf{c}_i$, and indicates that spurious minima exist.

2. SEPARABLE MULTI-NORMAL PDF

Consider the normal pdf $\mathcal{K}(x, \mu, \sigma)$ of a random Gaussian variable with mean μ and standard deviation σ:

$$\mathcal{K}(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \ . \qquad (3)$$

In the next subsections, we shall define the concepts of multi-normality and separability of probability density functions.

2.1 K-normal Density

The pdf of the random variable U is said to be K-normal if p_U has K normal modes, i.e. if

Michel Verleysen is a Senior Research Associate of the Belgian National Fund for the Scientific Research (F.N.R.S.).

EUSIPCO 05 - 13th European Signal Processing Conference
Antalya (Turkey), 4-8 September 2005, pp. 1107:1-4

$$p_U(x) = \sum_{i=1}^{K} \gamma_i \mathcal{K}(x, \mu_i, \sigma_i) \ , \qquad (4)$$

where γ_i are positive scaling factors ensuring that p_U integrates to one ($\sum_{i=1}^{K} \gamma_i = 1$). It is further assumed that $\mu_i \neq \mu_j$ if $i \neq j$, and that $\mu_1 < \mu_2 < \ldots < \mu_K$, without loss of generality.

2.2 Separability

The support of a one-dimensional Gaussian variable is $\mathbb{R}$. However, one can define a finite approximation Ω_i of this support. Indeed, the contribution of $\mathcal{K}(x, \mu_i, \sigma_i)$ 'far from the mean' is negligible, and if the support Ω_i is centered on the mean and large enough, we have

$$\int_{\Omega_i} \mathcal{K}(x, \mu_i, \sigma_i) \lessapprox 1 \ . \qquad (5)$$

In the following, a random variable U is said to be *K-normal separable* if the support of its pdf p_U can be approximated by the union of K finite disjoint intervals Ω_i, i.e. if for all $1 \leq i \leq K$ we have

$$\int_{\Omega_i} p_U(u)du \simeq \gamma_i \int_{\Omega_i} \mathcal{K}(x, \mu_i, \sigma_i) \simeq \gamma_i \ . \qquad (6)$$

3. ENTROPY OF A SEPARABLE MULTI-NORMAL DENSITY

3.1 Entropy estimator

If p_U is a K-normal separable distribution, we have:

$$\begin{aligned}
H(U) &= -\int_{-\infty}^{+\infty} \sum_{i=1}^{K} \gamma_i \mathcal{K}(x, \mu_i, \sigma_i) \log \Big\{ \sum_{i=1}^{K} \gamma_i \mathcal{K}(x, \mu_i, \sigma_i) \Big\} \\
&\overset{(a)}{\simeq} -\sum_{j=1}^{K} \int_{\Omega_j} \sum_{i=1}^{K} \gamma_i \mathcal{K}(x, \mu_i, \sigma_i) \log \Big\{ \sum_{i=1}^{K} \gamma_i \mathcal{K}(x, \mu_i, \sigma_i) \Big\} \\
&\overset{(b)}{=} -\sum_{i=1}^{K} \gamma_i \int_{\Omega_i} \mathcal{K}(x, \mu_i, \sigma_i) \log \gamma_i \mathcal{K}(x, \mu_i, \sigma_i) \\
&= -\sum_{i=1}^{K} \gamma_i \int_{\Omega_i} \mathcal{K}(x, \mu_i, \sigma_i) \Big\{ \log \gamma_i + \log \mathcal{K}(x, \mu_i, \sigma_i) \Big\} \\
&\overset{(c)}{\simeq} -\sum_{i=1}^{K} \gamma_i \Big\{ \log \gamma_i + \int_{\Omega_i} \mathcal{K}(x, \mu_i, \sigma_i) \log \mathcal{K}(x, \mu_i, \sigma_i) \Big\} \\
&\overset{(d)}{\simeq} -\sum_{i=1}^{K} \gamma_i \Big\{ \log \gamma_i + \underbrace{\int_{\mathbb{R}} \mathcal{K}(x, \mu_i, \sigma_i) \log \mathcal{K}(x, \mu_i, \sigma_i)}_{=-H(G_i)} \Big\} \\
&= \sum_{i=1}^{K} \gamma_i H(G_i) - \underbrace{\sum_{i=1}^{K} \gamma_i \log \gamma_i}_{\triangleq -H^{\gamma}} \ . \qquad (7)
\end{aligned}$$

In the previous development, (a) results from the fact that far from the μ_i's, $p_U \simeq 0$ since p_U is separable and $0 \log 0 = 0$ by convention. Relation (b) comes from the separability of p_U: in Ω_i, we can neglect the contribution of the $\mathcal{K}(x, \mu_j, \sigma_j)$ modes with respect to the $\mathcal{K}(x, \mu_i, \sigma_i)$ one, if $i \neq j$. In other words, in Ω_i, p_U is mainly determined by the i^{th} mode. If Ω_i is large enough, then (6) holds, leading to (c) and (d).

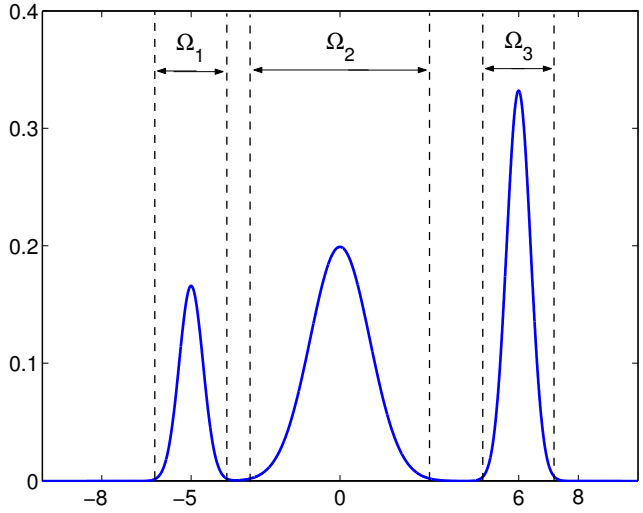


Figure 1: 3-normal separable pdf.

In equation (7), $H(G_i)$ is the entropy of a Gaussian variable G_i with $E\{G_i\} = \mu_i$ and $E\{G_i\}^2 - E\{G_i^2\} = \sigma_i^2$:

$$H(G_i) = \log \sqrt{2\pi\,\mathrm{e}} + \log \sigma_i \ . \qquad (8)$$

Hence, when approximation (7) holds, we will use the following approximation of a multi-normal separable random variable entropy:

$$\hat{H}(U) \triangleq \log \sqrt{2\pi\,\mathrm{e}} + \sum_{i=1}^{K} \gamma_i \log \sigma_i + H^{\gamma} \ . \qquad (9)$$

The relative error $\rho(U)$ resulting from the above approximation is defined by

$$\rho(U) \triangleq \left| \frac{H(U) - \hat{H}(U)}{H(U)} \right| \ . \qquad (10)$$

3.2 Example

In order to prove the efficiency of this estimator on multi-normal separable variable, $H(U)$ is compared to $\hat{H}(U)$ on the distribution given in Figure 1. We can assume that p_U is separable, according to the definition given in Section 2. The parameters of this 3-normal pdf are:

$$\left\{ \begin{array}{lcl} \mu & = & [-5,\ 0,\ 6] \\ \sigma & = & [2/5,\ 1,\ 2/5] \\ \gamma & = & [1/6,\ 1/2,\ 1/3] \end{array} \right. \ . \qquad (11)$$

In order to evaluate $H(U)$, the integral in the entropy definition is replaced by a Riemanian sum: $H(U) = -\sum_{u_i} p_U(u_i) \log p_U(u_i)$, where u_i ranges from -10 to 10 by increasing steps of 5.10^{-3}. The approximation $\hat{H}(U)$ has been computed through (11) and (9). In this example, we find $\rho(U) = 0.02\%$, which confirms the validity of approximator (9) for K-normal separable pdf.

i	$\mu_1(s_i)$	$\mu_2(s_i)$	$\gamma_1(s_i)$	$\gamma_2(s_i)$	$\sigma_1(s_i) = \sigma_2(s_i)$
1	-0.995	0.995	1/2	1/2	0.1
2	-0.81	1.22	3/5	2/5	0.1

Table 1: Parameters of two bimodal pdf.

4. ORTHOGONAL CONSTRAINT AND STIEFEL MANIFOLD

Let us focus on the simple case where two whitened sources have to be separated from two whitened mixtures. The system can be rewritten as :

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \underbrace{\begin{bmatrix} \sin\theta & \cos\theta \\ -\cos\theta & \sin\theta \end{bmatrix}}_{\mathbf{C}_\theta} \cdot \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} . \quad (12)$$

Note that since the whiteness property is invariant to any orthogonal transformation, we have that the output signals are whitened.

The set of $\mathbf{C}_\theta$ matrices forms a 1-D Stiefel manifold in the $\mathbb{R}^4$-space, since $\mathbf{C}_\theta \mathbf{C}_\theta^T = \mathbf{I}_2$ and θ is its only degree of freedom.

In the following, we will focus on the extraction of y_1 (the other source is known once θ is known). For clarity, we will adopt the following notation:

$$Z_\theta \triangleq y_1 = \sin\theta s_1 + \cos\theta s_2 \ . \quad (13)$$

Since both sources can be recovered for a specific $\theta \in [0, \pi/2]$, θ is constrained to be in the first quadrant, without loss of generality. In the next section, it is proven that $H(Z_\theta)$ may have mixing minima (i.e. for $\theta \notin \{0, \pi/2\}$), when s_1 and s_2 have specific probability density functions.

5. SPURIOUS MINIMA IN THE OUTPUT MARGINAL ENTROPY

Consider the bimodal separable pdf with parameters given in Table 1. They correspond to mutually independent zero-mean unit-variance source signals.

The entropies of these signal are: $H(s_1) = -0.27$ and $H(s_2) = -0.30$. Using the entropy estimator given by (9), we have $\rho(s_1), \rho(s_2) < 10^{-12}$.

Let us remark that the number of modes of p_{Z_θ} varies with θ, as is illustrated in Figure 2. Furthermore, for several θ, the modes of p_{Z_θ} are Gaussian-shape with small overlap; i.e. p_{Z_θ} is multi-normal separable. This is due to convolution properties of Gaussian functions. With a slight abuse of notation, let us denote by $\mu_i(\theta), \gamma_i(\theta)$ and $\sigma_i(\theta)$ the parameters of the i-th mode of the output pdf p_{Z_θ} (i.e. its mean, weight and variance).

The $\mu_i(\theta)$, $\gamma_i(\theta)$ and $\sigma_i(\theta)$ parameters can be computed. For instance, the trimodal case is obtained for $\theta_2 \triangleq \arctan \frac{\mu_2(s_2) - \mu_1(s_2)}{\mu_2(s_1) - \mu_1(s_1)}$ ($\gamma_1(\theta_2) = 0.3$, $\gamma_2(\theta_2) = 0.5$, $\gamma_3(\theta_2) = 0.2$ and $\sigma_{1,2,3}(\theta_2) = 0.1$), and it also exists two angles, noted θ_1 and θ_3, for which p_{Z_θ} is four-modal separable and such that $0 < \theta_1 < \theta_2 < \theta_3 < \pi/2$. For example, we can take $\theta_3 = \frac{13\pi}{36}$ and we have $\gamma_1(\theta_3) = \gamma_3(\theta_3) = 0.3$, $\gamma_2(\theta_3) = \gamma_4(\theta_3) = 0.2$ and $\sigma_{1,2,3,4}(\theta_3) = 0.1$ (see Figure 2). On the other hand, we can choose $\theta_1 = \pi/2 - \theta_3$ ($\sigma_i(\theta_1)$ and $\gamma_i(\theta_1)$ are the same as $\sigma_i(\theta_3)$ and $\gamma_i(\theta_3)$, except that the values of $\gamma_2(.)$ and $\gamma_3(.)$ have to be permuted). Note that the mean of the modes do not matter as soon as the modes have a negligible overlap.

The key point is that $\hat{H}(Z_0 = s_2) < \hat{H}(Z_{\pi/2} = s_1) < \hat{H}(Z_{\theta_2}) < \hat{H}(Z_{\theta_1}) \simeq \hat{H}(Z_{\theta_3})$. Indeed, $\hat{H}(Z_{\theta_2}) = 0.21$ while $\hat{H}(Z_{\theta_1}) = 0.70$. In addition, p_{Z_θ} is multi-normal separable for $\theta \in \{0, \theta_1, \theta_2, \theta_3, \pi/2\}$ and therefore, according to Section 3.2, the approximator (9) is valid: we must have $H(Z_\theta) \simeq \hat{H}(Z_\theta)$. As a consequence, $H(Z_\theta)$ must have a mixing minimum for θ in (θ_1, θ_3). This result can be observed in Figure 3 where $H(Z_\theta)$ is plotted w.r.t. θ in the first quadrant. Note that p_{Z_θ} has been computed by convoluting $p_{\sin\theta s_1}$ and $p_{\cos\theta s_2}$. By doing so, $H(Z_\theta)$ cannot be numerically evaluated with a high precision for $\theta \simeq k\pi/2$; this is why $H(Z_\theta)$ is plotted for $\theta \in [\epsilon, \pi/2 - \epsilon]$, where ϵ is a small positive number. Nevertheless, it is obvious that one must have $H(Z_\theta) \to H(Z_{k\pi/2})$ when $\theta \to k\pi/2$.

6. CONCLUSION

In this paper, the concept of separability of a multi-normal density is presented. A simple approximator is derived to estimate the entropy of a variable having such density. Next, we have focused on the separation of two sources having bimodal densities. The Gaussian mode variances of the two pdf have been chosen small enough with respect to the intermodal distance. The major consequence of that choice is that the whitened mixture of these sources may have 2, 3 or 4 Gaussian modes, depending on the mixture coefficients. Then, the estimator has been used to estimate the entropy of the output signal, when the mixture weights are such that the output density is separable (i.e. for densities for which the estimator is valid). Mathematical results show that in the analyzed case, the marginal entropic contrast has mixing minima under the whitening constraint.

The existence of spurious minima in the entropy cost function for source separation should be taken into account in all gradient-based algorithms that rely on this criterion; otherwise, spurious solutions to the BSS problem could be obtained. In a more general framework, the above results can be used to estimate the entropy of a linear mixture of multi-normal separable sources when the mixture weights are such that the mixture pdf is also multi-normal separable.

REFERENCES

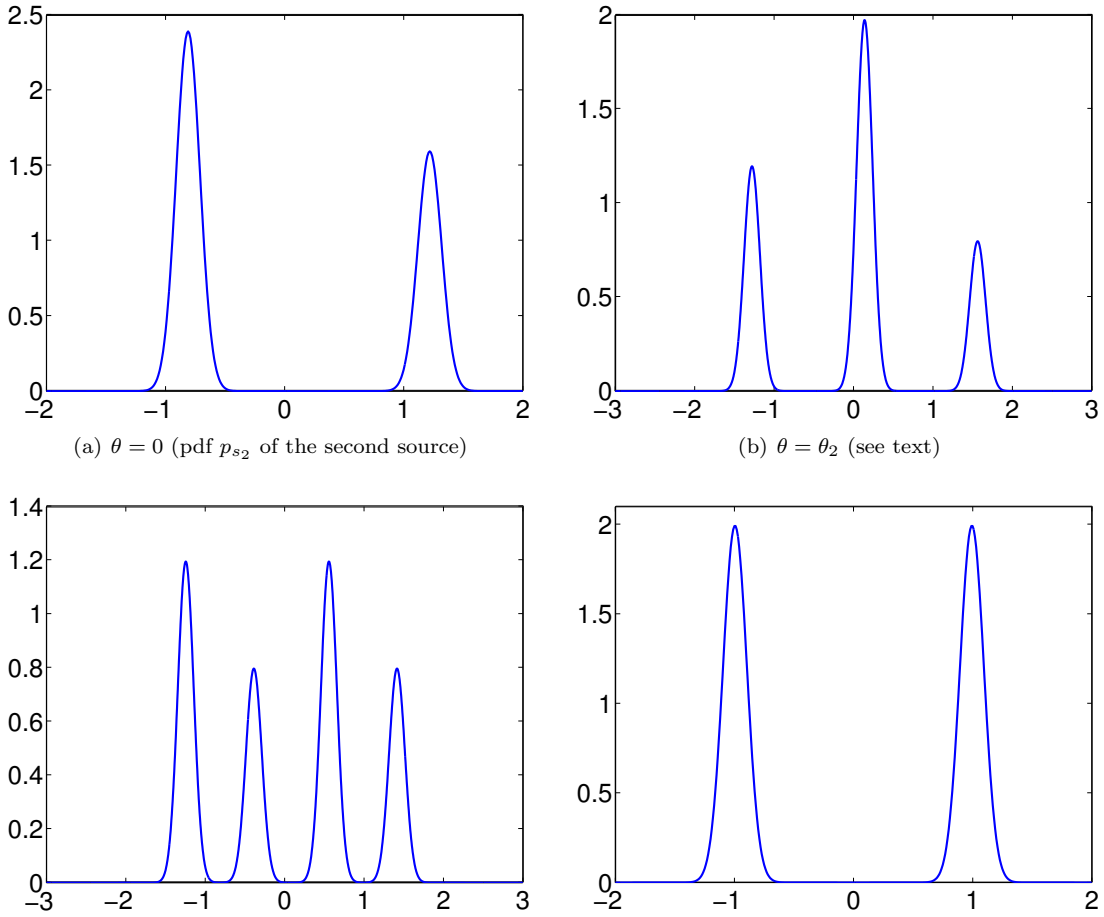
[1] F. Vrins and M. Verleysen. On the entropy minimization of a linear mixture of variables for source separation. *Signal Processing*, 85(5):1029–1044, 2005.

[2] S. Cruces, A. Cichocki, and S. Amari. The minimum entropy and cumulants based contrast functions for blind source extraction. In *proc. of IWANN 2001*, LNCS, pages 786–793. October 2001.

[3] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley and sons, 1991.

[4] S. Haykin, editor. *Unsupervised Adaptive Filtering vol.1 : Blind Source Separation (ch. IV, pp 171-173)*. John Willey ans Sons, Inc., New York, 2000.

EUSIPCO 05 - 13th European Signal Processing Conference
Antalya (Turkey), 4-8 September 2005, pp. 1107:1-4



(a) $\theta = 0$ (pdf p_{s_2} of the second source)

(b) $\theta = \theta_2$ (see text)

(c) $\theta = \frac{13\pi}{36}$ (example of an acceptable value for θ_3)

(d) $\theta = \frac{\pi}{2}$ (pdf p_{s_1} of the first source)

Figure 2: PDF p_{Z_θ} for several values of $\theta \in [0, \pi/2]$.

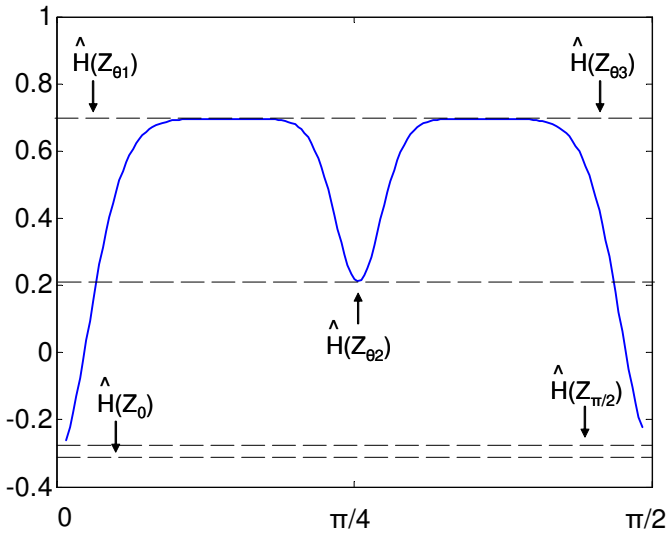


Figure 3: Evolution of $H(Z_\theta)$ for $\theta \in [\epsilon, \pi/2 - \epsilon]$, with $\epsilon \simeq 0.03$. For θ corresponding to a K-modal separable p_{Z_θ}, $H(Z_\theta) \simeq \hat{H}(Z_\theta)$. This approximation is reliable for $\theta \in \{0, [0.4, 0.55], \theta_2, [1.02, 1.17], \frac{\pi}{2}\}$