Local minima of information-theoretic criteria in blind source separation

Dinh-Tuan Pham, *Member, IEEE* and Frédéric Vrins

***Abstract*—Recent simulation results have indicated that spurious minima in information-theoretic contrasts with orthogonality constraint for blind source separation may exist. Nevertheless, those results involve approximations (e.g. density estimation), so that they do not constitute an absolute proof. In this letter, the problem is tackled from a theoretical point of view: an example is provided for which it is rigorously proven that spurious minima can exist in both mutual information and negentropy contrasts. The proof is based on a Taylor expansion of the entropy function.**

***Index Terms*—Blind source separation. Independent Component Analysis. Entropy. Mutual Information**

EDICS Category: SAS-ICAB

I. Introduction

Blind source separation aims to extract source vector $\mathbf{S} = [S_1 \quad \cdots \quad S_M]^{\mathrm{T}}$ from observed mixtures $\mathbf{X} = [X_1 \quad \cdots \quad X_N]^{\mathrm{T}}$, relying mainly only on the independence assumption of the sources [3]. The simplest and most widely used mixture model assumes noise-free linear mixtures with $K \triangleq M = N$: $\mathbf{X} = \mathbf{AS}$ where $\mathbf{A}$ is a $K \times K$ invertible matrix. As no specific knowledge on the distribution of the S_k and the mixing matrix $\mathbf{A}$ is available, a common separation method is to find an unmixing matrix $\mathbf{B}$ such that the extracted sources, which are the components $Y_1, \dots, Y_K$ of

$$\mathbf{Y} = \mathbf{BX} \tag{1}$$

be as independent as possible. Another method, which exploits the link between independence and non Gaussianity, extracts the most non Gaussian source by finding the vector $\mathbf{b}$ such that $\mathbf{b}^{\mathrm{T}}\mathbf{X}$ is most non Gaussian. Once this source is extracted, the second most non Gaussian source can be extracted in the same way but constraining the vector $\mathbf{b}$ such that $\mathbf{b}^{\mathrm{T}}\mathbf{X}$ is uncorrelated with the previously extracted sources, and so on ... This is the underlying principle of the deflation approach of the fastICA algorithm [8].

The above methods require a measure of dependence and of non Gaussianity. A popular measure of statistical dependence is the mutual information [3], [4]. On the other hand, a statistically efficient measure of non Gaussianity is the negentropy [9]. The two measures are related, as they both involve the entropy of the extracted sources. Although methods based on them would find the correct sources if the *absolute minimum* of the criteria is reached, in practice, one may get stuck in local minima. Experimental investigations [6], [7], in the case of information-theoretic contrasts with orthogonality constraint reveal the existence of local minima of the criterion, which *do not correspond to a source but to a mixture*. In this paper, we shall provide a specific example for which the existence of such spurious (local) minima can be theoretically rigorously proved, both for the mutual information and the negentropy-based criterion.

II. Spurious minima of the mutual information

We consider the separation method which consists in minimizing the mutual information criterion $I(Y_1, \dots, Y_K)$ [2] between the reconstructed sources $Y_1, \dots, Y_K$. Mutual information can be written in term of entropies as $I(Y_1, \dots, Y_K) = \sum_{k=1}^{K} H(Y_k) - H(\mathbf{Y})$ where $H(Y_k) = -\int p_{Y_k}(y) \log p_{Y_k}(y)\; dy$ is the entropy of Y_k and $H(\mathbf{Y}) \triangleq H(Y_1, \dots, Y_K)$ is defined similarly [2]. As $\mathbf{Y} = \mathbf{BX}$, it can be seen (see e.g. [4]) that $H(\mathbf{Y}) = H(\mathbf{X}) + \log|\det \mathbf{B}|$. Therefore, minimizing $I(Y_1, \dots, Y_K)$ is equivalent to minimizing

$$C(\mathbf{B}) = \sum_{k=1}^{K} H(Y_k) - \log\det \mathbf{B} = I(Y_1, \dots, Y_K) - H(\mathbf{X}).$$

A. Expansion of the criterion up to second order

To see if a point $\mathbf{B}$ minimizes locally C, we perform a Taylor expansion of C around $\mathbf{B}$ up to second order. Because of the multiplicative structure of the mixture model, it is of interest to consider the relative (rather than absolute) increment of the parameter $\mathbf{B}$. More precisely, we make a Taylor development of $C(\mathbf{B} + \mathcal{E}\mathbf{B})$ up to second order with respect to a "small matrix" $\mathcal{E}$. Using the result of [5], we have, putting $\delta Y_i = \sum_{k=1}^{K} \mathcal{E}_{ik} Y_k$:

$$H(Y_i + \delta Y_i) = H(Y_i) + \mathrm{E}[\psi_{Y_i}(Y_i)\delta Y_i] + \frac{1}{2}\{\mathrm{E}[\mathrm{var}(\delta Y_i|Y_i)\psi'_{Y_i}(Y_i)] - [\mathrm{E}(\delta Y_i|Y_i)]'^2\} + o(\|\mathcal{E}\|^2)$$

where $\psi_{Y_i} = (-\log p_{Y_i})'$ is the score function of Y_i, $'$ denoting the derivative, and $\mathrm{E}(\cdot|Y_i)$ denotes the conditional expectation given Y_i and $\mathrm{var}(\delta Y_i|Y_i) = \mathrm{E}[(\delta Y_i)^2|Y_i] - \mathrm{E}(\delta Y_i|Y_i)^2$ is the conditional variance of δY_i given Y_i. Further, $\log\det(\mathbf{B} + \mathcal{E}\mathbf{B}) = \log\det \mathbf{B} + \mathrm{tr}(\mathcal{E}) - \frac{1}{2}\mathrm{tr}(\mathcal{E}^2) + o(\|\mathcal{E}\|^2)$. Therefore, noting that $\mathrm{E}[\psi_{Y_k}(Y_k)Y_k] = 1$ by integration by parts (see [4]),

$$C(\mathbf{B} + \mathcal{E}\mathbf{B}) = C(\mathbf{B}) + \sum_{i \neq j} \mathrm{E}[\psi_{Y_i}(Y_i)Y_j]\mathcal{E}_{ij} + \frac{1}{2}\sum_{i,j,k}\{\mathrm{E}[\mathrm{cov}(Y_j, Y_k|Y_i)\psi'_{Y_i}(Y_i)] - [\mathrm{E}(Y_j|Y_i)]'[\mathrm{E}(Y_k|Y_i)]'\} \times \mathcal{E}_{ij}\mathcal{E}_{ik} + \frac{1}{2}\sum_{i,j}\mathcal{E}_{ij}\mathcal{E}_{ji} + o(\|\mathcal{E}\|^2)$$

Dinh-Tuan Pham is with Laboratoire de Modélisation et Calcul, CNRS, BP 53, 38041 Grenoble, France. E-mail: Dinh-Tuan.Pham@imag.fr, Phone: +33 476514423, Fax: +33 476631263

Frédéric Vrins is with the UCL Machine Learning Group, Université catholique de Louvain, Place du Levant 3, 1348 Louvain-la-Neuve (Belgium). E-mail: vrins@dice.ucl.ac.be, URL: www.ucl.ac.be/mlg, Phone: +32 10 47.81.33, Fax: +32 10 47.25.98

where $\mathrm{cov}[Y_j, Y_k|Y_i] = \mathrm{E}[Y_j Y_k|Y_i] - \mathrm{E}[Y_j|Y_i]\mathrm{E}[Y_k|Y_i]$.

The above expansion shows that $\mathbf{B}$ is a stationary point of C if $\mathrm{E}[\psi_{Y_j}(Y_j)Y_k] = 0$ for $j \neq k$. To see if $\mathbf{B}$ is indeed a local minimum, one has to look at the second order term in the above expansion, which is quite involved. Therefore, we shall focus on the case of two sources.

B. Existence of spurious minima in the case of two sources

Consider the case of two sources ($K = 2$) with the same density function p_S. Since $I(Y_1, Y_2) \geq 0$ with equality if and only if the variables Y_k are mutually independent, C admits a global minimum at $\mathbf{B} = \mathbf{PDA}^{-1}$ where $\mathbf{P}$ and $\mathbf{D}$ are permutation and diagonal matrices, respectively [1]. We now show that for certain source distribution,

$$\mathbf{B} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \mathbf{A}^{-1} \tag{2}$$

is also a local minimum of C. This is also true if the above right hand side is left multiplied by $\mathbf{PD}$, since C is invariant when one left multiplies its argument by any permutation or diagonal matrix.

For $\mathbf{B}$ given in (2), $Y_1 = S_1 + S_2$, $Y_2 = S_2 - S_1$. Therefore

$$\mathrm{E}[\psi_{Y_1}(Y_1)Y_2] = \mathrm{E}[\psi_{Y_1}(S_1+S_2)S_2] - \mathrm{E}[\psi_{Y_1}(S_1+S_2)S_1],$$

$$\mathrm{E}[\psi_{Y_2}(Y_2)Y_1] = \mathrm{E}[\psi_{Y_2}(S_2-S_1)S_1] + \mathrm{E}[\psi_{Y_2}(S_2-S_1)S_2].$$

But since the joint distribution of (S_1, S_2) is the same as that of (S_2, S_1), one can permute S_1 and S_2 in the above right hand sides without changing their value. Hence these right hand sides vanish, noting that $Y_2 = S_2 - S_1$ has the same distribution as $-Y_2$ and therefore ψ_{Y_2} is an odd function.

The above results show that the matrix $\mathbf{B}$ in (2) is a stationary point of C. To see if it is a local minimum point, we consider the expansion of $C(\mathbf{B} + \mathcal{E}\mathbf{B})$ up to second order. Again, since one can permute S_1 and S_2 without changing their joint distributions, $\mathrm{E}[S_2|S_1 + S_2] = \mathrm{E}[S_1|S_2 + S_1]$ and hence $\mathrm{E}[Y_2|Y_1] = 0$. In the case where p_S is symmetric so that $-S_1$ has the same distribution as S_1, by the same argument as before with S_1 replaced by $-S_1$: $\mathrm{E}[S_2|S_2 - S_1] = -\mathrm{E}[S_1|S_2 - S_1]$ and hence $\mathrm{E}[Y_1|Y_2] = 0$. Therefore from the result of II-A and noting that $\mathrm{E}[\mathrm{E}[Y_j^2|Y_i]\psi'_{Y_i}(Y_i)] = \mathrm{E}[Y_j^2\psi'_{Y_i}(Y_i)]$ [10]:

$$C(\mathbf{B} + \mathcal{E}\mathbf{B}) = C(\mathbf{B}) + \frac{1}{2}\{\mathrm{E}[Y_2^2\psi'_{Y_1}(Y_1)]\mathcal{E}_{12}^2 + \mathrm{E}[Y_1^2\psi'_{Y_2}(Y_2)]\mathcal{E}_{21}^2\} + \mathcal{E}_{12}\mathcal{E}_{21} + o(\|\mathcal{E}\|^2).$$

The above expansion shows that $\mathbf{B}$ is a local minimum point if and only if $\mathrm{E}[Y_2^2\psi'_{Y_1}(Y_1)]\mathrm{E}[Y_1^2\psi'_{Y_2}(Y_2)] > 1$. But since the joint distribution of Y_1, Y_2 is the same as the one of Y_2, Y_1, this condition is equivalent to $\mathrm{E}[Y_2^2\psi'_{Y_1}(Y_1)] > 1$.

Inspired by simulations presented in [7], we now show that the above condition is satisfied (so that $\mathbf{B}$ is a spurious local minimum) for a source density being a mixture of two normal densities

$$p_S(s) = \{\phi[(s+1)/\sigma] + \phi[(s-1)/\sigma]\}/2$$

with sufficiently small σ (hence is bimodal), where $\phi(s) = \exp(-s^2/2)/\sqrt{2\pi}$ is the standard normal density.

Lemma 1: Let S_1 and S_2 have the same density p_S. Then $Y_1 = S_1 + S_2$ and $Y_2 = S_2 - S_1$ have the same density

$$p_Y(y) = \frac{1}{4}\phi\Big(\frac{y+2}{\sqrt{2}\sigma}\Big) + \frac{1}{2}\phi\Big(\frac{y}{\sqrt{2}\sigma}\Big) + \frac{1}{4}\phi\Big(\frac{y-2}{\sqrt{2}\sigma}\Big).$$

Their common score function ψ_Y admits the derivative

$$\psi'_Y(y) = \frac{1}{2\sigma^2} - \frac{w_{-1}(y)w_0(y) + w_1(y)w_0(y) + 4w_{-1}(y)w_1(y)}{\sigma^4}$$

where

$$w_0(y) = \frac{2\phi(y/\sqrt{2}\sigma)}{\phi[(y+2)/\sqrt{2}\sigma] + 2\phi(y/\sqrt{2}\sigma) + \phi[(y-2)/\sqrt{2}\sigma]}$$

$$w_{\mp 1}(y) = \frac{\phi[(y \pm 2)/\sqrt{2}\sigma]}{\phi[(y+2)/\sqrt{2}\sigma] + 2\phi(y/\sqrt{2}\sigma) + \phi[(y-2)/\sqrt{2}\sigma]}$$

Further, $\mathrm{E}(Y_2^2|Y_1 = y) = 2\sigma^2 + 4w_0(y)$.

Lemma 2: The expectation $\mathrm{E}[Y_2^2\psi'_{Y_1}(Y_1)]$ equals

$$1 + \frac{1}{\sigma^2} - \int \frac{[\sigma^2 + 2w_0(y)][w_0(y) + 2w_1(y)]}{\sigma^4}\phi\Big(\frac{y+2}{\sqrt{2}\sigma}\Big)dy.$$

The last term in the above expression tends to 0 as $\sigma \to 0$ and hence $\mathrm{E}[Y_2^2\psi'_{Y_1}(Y_1)] \to \infty$ as $\sigma \to 0$.

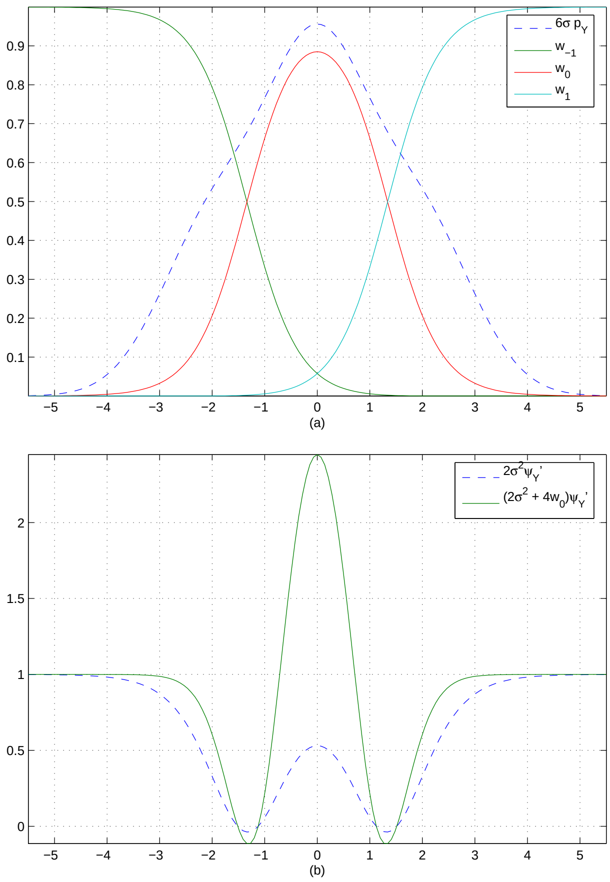

Fig. 1. Plot of $6\sigma p_Y$ and w_{-1}, w_0, w_1 (a), and $2\sigma^2\psi'_Y$ and $(2\sigma^2+4w_0)\psi'_Y$ (b). Density p_Y is unimodal ($\sigma = 0.7$): spurious minima of mutual information and entropy cannot be observed (see text).

The above results show that for σ small enough, there is a spurious minima at the point (2). Figure 1 illustrates the case $\sigma = 0.7$ for which $\mathrm{E}[Y_2^2\psi'_{Y_1}(Y_1)] = 0.9489 < 1$ and Figure 2

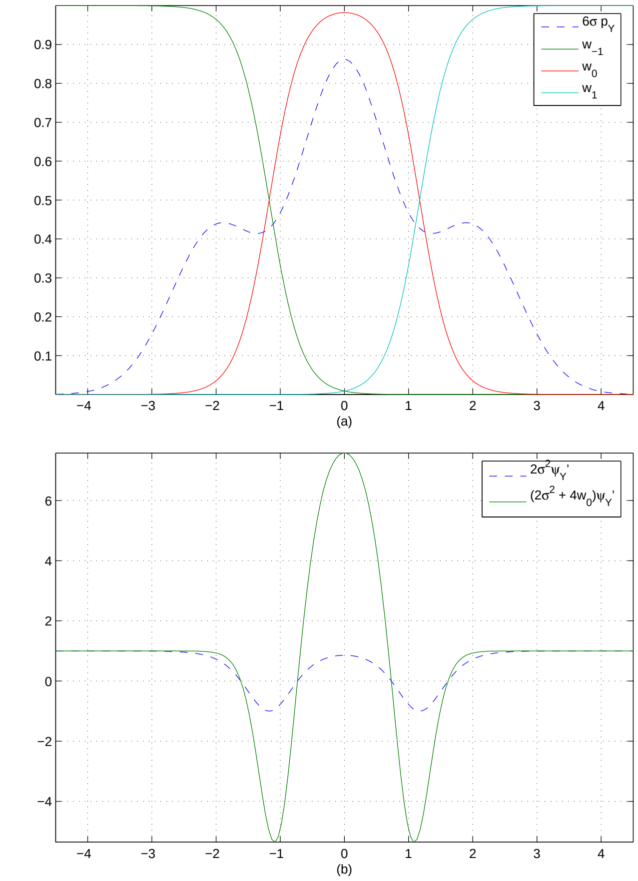

Fig. 2. Plot of $6\sigma p_Y$ and w_{-1}, w_0, w_1 (a), and $2\sigma^2\psi'_Y$ and $(2\sigma^2 + 4w_0)\psi'_Y$ (b). Distribution p_Y is trimodal ($\sigma = 0.5$): spurious minima of mutual information and entropy can be observed (see text).

illustrates the case $\sigma = 0.5$ for which $\mathrm{E}[Y_2^2\psi'_{Y_1}(Y_1)] = 1.5412 > 1$. One can see from figures 1(a) and 2(a) that as σ decreases, p_Y changes from an unimodal to a trimodal structure and the w_i approach a step function, and from figures 1(b) and 2(b) that $2\sigma^2\psi'_Y$ approaches 1 inside the three regions $(-\infty, -1), (-1, 1), (1, \infty)$, with "dips" at the transition points (the "dips" being sharper for smaller σ). The product of ψ'_Y with the function $y \mapsto \mathrm{E}[Y_2^2|Y_1 = y]$, which equals $2\sigma^2 + 4w_0$, produces a curve of similar shape as $2\sigma^2\psi'_Y$, but with a higher level in the central region due to the term $4w_0$. Its integral with respect to the density p_Y yields $\mathrm{E}[Y_2^2\psi'_{Y_1}(Y_1)]$. As p_Y is low in the neighborhood of the transition points ± 1, even more so as σ becomes smaller, the effect of the "dips" is attenuated and the integral should become larger as σ decreases, since the function takes a higher value inside the central region. This explains why one gets a larger value of $\mathrm{E}[Y_2^2\psi_{Y_1}(Y_1)]$. Figure 3 plots this quantity versus σ. One can see that when σ decreases beyond the value 0.63 (approximately) this quantity becomes greater than 1.

III. Spurious maxima of the negentropy

We consider the negentropy based FasICA, which consists in maximizing the negentropy of $\mathbf{b}^{\mathrm{T}}\mathbf{X}$. Recall that the negentropy of a random variable Y is the difference between the entropy of a Gaussian random variable of the same variance

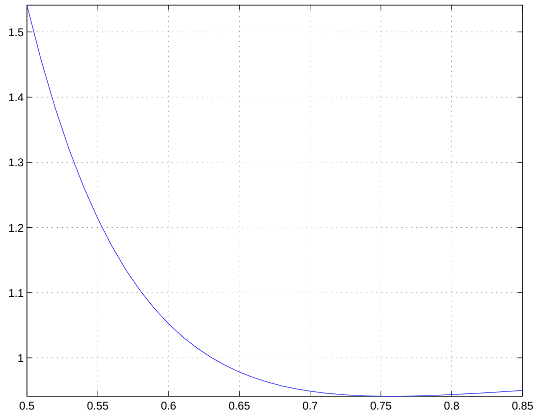

Fig. 3. Plot of $\mathrm{E}[Y_2^2\psi'_{Y_1}(Y_1)]$ versus σ.

as Y and the entropy of Y, that is $\frac{1}{2}\log[2\pi e\mathrm{var}(Y)] - H(Y)$ where $\mathrm{var}(\cdot)$ denotes the variance and $e = \exp(1)$. One may assume that the S_k have a same variance, since one can divide any of them by a constant and multiplies the corresponding columns of $\mathbf{A}$ by the same constant, without changing the mixture model. Since the negentropy is scale invariant, one may normalize $\mathbf{b}$ such that $\mathbf{b}^T\mathbf{X}$ has the same variance. Thus, putting $\mathbf{w} = \mathbf{A}^{\mathrm{T}}\mathbf{b}$, negentropy based FastICA amounts to minimizing $H(\mathbf{w}^{\mathrm{T}}\mathbf{S})$ under the constraint $\|\mathbf{w}\| = 1$.

Consider the case of two sources, one may parameterize $\mathbf{w}$ as $[\cos\theta \quad \sin\theta]^{\mathrm{T}}$. Put $Z_\theta = \mathbf{w}^{\mathrm{T}}\mathbf{S} = \cos\theta S_1 + \sin\theta S_2$, it is easy to see that a small change $\delta\theta$ in θ induces a change

$$\delta Z_\theta = (-\sin\theta S_1 + \cos\theta S_2)\delta\theta - \frac{1}{2}Z_\theta(\delta\theta)^2$$

in Z_θ up to second order in $\delta\theta$. Thus by the same calculation as in section II, using the result of [5], one gets, putting $Z_\theta^\perp = (-\sin\theta S_1 + \cos\theta S_2)$ and noting that $\mathrm{E}[\psi_{Z_\theta}(Z_\theta)Z_\theta] = 1$,

$$H(Z_\theta + \delta Z_\theta) \approx H(Z_\theta) + \mathrm{E}[\psi_{Z_\theta}(Z_\theta)Z_\theta^\perp]\delta\theta + \frac{1}{2}\{\mathrm{E}[\mathrm{var}(Z_\theta^\perp|Z_\theta)\psi'_{Z_\theta}(Z_\theta)] - [\mathrm{E}(Z_\theta^\perp|Z_\theta)]'^2 - 1\}(\delta\theta)^2$$

up to second order in $\delta\theta$.

The above result shows that a stationary point of $H(Z_\theta)$ (as a function of θ) occurs when $\mathrm{E}[\psi_{Z_\theta}(Z_\theta)Z_\theta^\perp] = 0$. Clearly, this is achieved for $\theta = 0$ and $\theta = \pi/2$ since $Z_0 = -Z_{\pi/2}^\perp = S_1$, $Z_{\pi/2} = Z_0^\perp = S_2$ and S_1 and S_2 aare independent. The points $\theta = 0$ and $\theta = \pi/2$ are actually local minima of $H(Z_\theta)$ if p_S is non Gaussian. Indeed, the second derivative of $H(Z_\theta)$ at $\theta = 0$ and $\theta = \pi/2$ reduces to $\mathrm{var}(S_2)\mathrm{E}[\psi'_{S_1}(S_1)] - 1$ and $\mathrm{var}(S_1)\mathrm{E}[\psi'_{S_2}(S_2)] - 1$ respectively. But for any random variable Y, $\mathrm{E}[\psi'_Y(Y)] = \mathrm{E}[\psi_Y^2(Y)]$ by integration by parts and $\mathrm{var}(Y)\mathrm{E}[\psi_Y^2(Y)] \geq 1$ by the Schwartz inequality, noting that $\mathrm{E}[\psi_Y(Y)Y] = 1$ and $\mathrm{E}[\psi_Y(Y)] = 0$. The inequality is strict unless ψ_Y is linear, that is Y is Gaussian. Note that since $H(Z_\theta)$ is periodic (with respect to θ) of period π, then function $H(Z_\theta)$ admits local minima for θ in $\{p\pi/2 | p \in \mathbb{Z}\}$.

Same arguments as in section II show that there are two other stationary points of $H(Z_\theta)$ at $\theta = \pi/4$, for which $Z_\theta = Y_1/\sqrt{2}$ and $Z_\theta^\perp = Y_2/\sqrt{2}$, and at $\theta = 3\pi/4$, for which $Z_\theta = -Y_2/\sqrt{2}$ and $Z_\theta^\perp = -Y_1/\sqrt{2}$. To see if they are the local minima, we look at the second derivative of $H(Z_\theta)$. As before, $\mathrm{E}[Y_2|Y_1] = 0$ and if p_S is symmetric, $\mathrm{E}[Y_1|Y_2] = 0$. In this case, the second derivative of $H(Z_\theta)$ at $\theta = \pi/4$ and $\theta = 3\pi/4$ reduces to $\mathrm{E}[Y_2^2\psi_{Y_1}(Y_1)] - 1$ and $\mathrm{E}[Y_1^2\psi_{Y_2}(Y_2)] - 1$, respectively, which are equal. Thus the condition for which these points to be local maxima of the negentropy are the same as the one for which the point (2) is a local minimum of the mutual information criterion.

IV. Conclusion

This letter gives an example of source density for which it can be proven that spurious minima of mutual information and marginal entropy exist in the BSS context. Our theoretical approach confirms the findings of earlier experimental studies based on an estimation of density and entropy through simulation. Although our result concern a specific example, the phenomenon of spurious minima seem to occurs generally for strongly multimodal source distribution.

Appendix: Proofs of Lemmas

Proof of Lemma 1

Note that (S_1, S_2) is distributed as $(U_1 + \sigma Z_1, U_2 + \sigma Z_2)$ where U_1, U_2 are independent Bernoulli variables taking the value ± 1 with probability $1/2$ and Z_1, Z_2 are independent standard normal variables independent of U_1, U_2. Thus (Y_1, Y_2) is distributed as $(U_1+U_2+\sqrt{2}\sigma Z_1', U_2-U_1+\sqrt{2}\sigma Z_2')$ where Z_1', Z_2' are also independent standard normal variables independent of U_1, U_2. Since $U_1 + U_2$ and $U_1 - U_2$ both take the value ± 2 with probability 1/4 and 0 with probability 1/2, Y_1 and Y_2 has the same distribution with density as p_Y given in the Lemma. Direct calculation yields $\psi_Y(y) = -p_Y'(y)/p_Y(y) = 1/(2\sigma^2)\sum_{i=-1}^{1}(y-2i)w_i(y)$ with w_i as defined in the Lemma. Noting that

$$w_i'(y) = \sum_{j=-1}^{1} \frac{y-2j}{2\sigma^2} w_i(y) w_j(y) - \frac{y-2i}{2\sigma^2} w_i(y)$$

and $1 - w_i = \sum_{j\neq i} w_j$, $(1/2\sigma^2)\sum_{i=-1}^{1}(y-2i)w_i'(y)$ equals

$$\sum_{-1\le i<j\le 1} \frac{w_i(y)w_j(y)}{4\sigma^4}[2(y-2i)(y-2j)-(y-2i)^2-(y-2j)^2]$$

$$= -\frac{1}{\sigma^4}\sum_{-1\le i<j\le 1}(j-i)^2 w_i(y) w_j(y).$$

This yields the expression for ψ_Y' given in the Lemma.

To compute $\mathrm{E}(Y_2^2|Y_1 = y)$, note that the conditional distribution of (U_1, U_2) given $Y_1 = y$ is

$$\begin{aligned}
P(U_1 = 1, U_2 = 1|Y_1 = y) &= w_1(y),\\
P(U_1 = 1, U_2 = -1|Y_1 = y) &= w_0(y)/2,\\
P(U_1 = -1, U_2 = 1|Y_1 = y) &= w_0(y)/2,\\
P(U_1 = -1, U_2 = -1|Y_1 = y) &= w_{-1}(y).
\end{aligned}$$

Therefore, conditionally on $Y_1 = y$, $Y_2 = U_2 - U_1 + \sqrt{2}\sigma Z_2'$ is distributed as $\sqrt{2}\sigma Z_2'$ with probability $w_{-1}(y) + w_1(y)$ and as $\sqrt{2}\sigma Z_2' \pm 2$ with probability $w_0(y)/2$ each. This yields the expression for $\mathrm{E}(Y_2^2|Y_1 = y)$ given in the Lemma. ∎

Proof of Lemma 2

We have $\mathrm{E}[Y_2^2\psi_{Y_1}'(Y_1)] = \int \mathrm{E}(Y_2^2|Y_1 = y)\psi_{Y_1}'(y)p_Y(y)dy$. Hence noting that $w_0(y) = \phi[y/(\sqrt{2}\sigma)]/[2p_Y(y)]$, $w_{\pm1}(y) = \phi[(y\mp 2)/(\sqrt{2}\sigma)]/[4p_Y(y)]$, and that w_0 is an even function, one get the first result of the Lemma.

We now derive an upper bound for $g(y) = [\sigma^2 + 2w_0(y)]$ $[w_0(y) + 2w_1(y)]/\sigma^4$. We have,

$$\frac{w_0(y)}{w_{-1}(y)} = 2\exp\Big[\frac{(y+2)^2 - y^2}{4\sigma^2}\Big] = 2\exp\Big(\frac{y+1}{\sigma^2}\Big).$$

Thus, since $w_{-1} = 1 - w_1 - w_0$

$$w_0(y) = \frac{2e^{(y+1)/\sigma^2}}{1 + 2e^{(y+1)/\sigma^2}}[1 - w_1(y)] \le 2e^{(y+1)/\sigma^2}.$$

Similarly, $w_0(y)/w_1(y) = 2\exp\{[(y-2)^2 - y^2]/(4\sigma^2)\} = 2\exp[(1-y)/\sigma^2]$ and since $w_0 = 1 - w_1 - w_{-1}$:

$$w_1(y) = \frac{1 - w_{-1}(y)}{1 + 2e^{(1-x)/\sigma^2}} \le \frac{1}{1 + 2e^{(1-x)/\sigma^2}} \le \frac{e^{(y-1)/\sigma^2}}{2}$$

Thus, for $y \le -1 - \xi$, one has

$$g(y) \le (1 + 4e^{-\xi/\sigma^2}/\sigma^2)[2e^{-\xi/\sigma^2} + e^{-(2-\xi)/\sigma^2}]/\sigma^2$$

If we choose $\xi = \xi(\sigma)$ such that $\xi/\sigma^2 + \log\sigma^2 \to \infty$ and $\xi \to 0$ as $\sigma \to 0$, then both $e^{-\xi/\sigma^2}/\sigma^2$ and $e^{-(2-\xi)/\sigma^2}/\sigma^2$ tend to 0 as $\sigma \to 0$. Hence

$$\int_{-\infty}^{-1-\xi} g(y)\phi\Big(\frac{y+2}{\sqrt{2}\sigma}\Big)dy \to 0$$

For $x \ge -1 - \xi$, one can bound $w_0(y)$ and $w_1(y)$ by 1, hence $g(y)$ by $2(\sigma^2 + 2)$. Therefore

$$\int_{-1-\xi}^{\infty} g(y)\phi\Big(\frac{y+2}{\sqrt{2}\sigma}\Big)dy \le \frac{3(\sigma^2+2)}{\sigma^4}\Big[1 - \Phi\Big(\frac{1-\xi}{\sqrt{2}\sigma}\Big)\Big]$$

where $\Phi(y) = \int_{-\infty}^{y} e^{-y^2/2}dy/\sqrt{2\pi}$ is the cumulative distribution function of the normal distribution. But we know that $1 - \Phi(y) \le e^{-y^2/2}/(y\sqrt{2\pi})$, hence

$$\frac{1}{\sigma^4}\Big[1 - \Phi\Big(\frac{1-\xi}{\sqrt{2}\sigma}\Big)\Big] \le \frac{\exp[-(1-\xi)^2/(4\sigma^2)]}{(1-\xi)\sqrt{\pi}\sigma^3} \to 0 \quad \text{as } \sigma \to 0,$$

since $\xi \to 0$ as $\sigma \to 0$. ∎

References

[1] Comon, P. "Independent component analysis – a new concept?," *Signal Processing*, vol. 36, pp. 287–314, 1994.

[2] Cover, Th. and Thomas, J. A. *Elements of Information Theory*. New-York: Wiley, 1991

[3] Cardoso, J.-F. "Blind Signal Separation: Statistical principles", *Proc. IEEE*, **90**, 2009–2026. Aug. 1998.

[4] Pham, D. T. "Blind Separation of Instantaneous Mixture of Sources via an Independent Component Analysis." *IEEE Trans. Signal Processing*, **44**, 11, 2768–2779, 1996.

[5] Pham, D. T. "Entropy of a Variable Slightly Contaminated With Another." *IEEE Signal Proc. Lett.*, To appear.

[6] Vrins, F. and Verleysen M. "On the Entropy Minimization of a Linear Mixture of Variables for Source Separation ." *Signal Process.*, In press.

[7] Vrins, F. and Verleysen M. "Information Theoretic versus Cumulant-based Contrasts for Multimodal Source Separation." *IEEE Signal Process. Lett.*, **12**, 3, 190–193, 2005.

[8] Hyvärinen, A. "Fast and Robust Fixed-Point Algorithms For Independent Component Analysis." *IEEE Trans. Neural Networks*, **10**, 3, 626–634, 1999.

[9] Hyvärinen, A. "Survey on independent component analysis." *Neural Computing Surveys*, **2**, 94–128, 1999.

[10] Gray, R. and Davisson, L. "An Introduction to Statistical Signal Processing." *Cambridge University Press*, p. 208, 2004.