On the Risk of Using Rényi's Entropy for Blind Source Separation

Dinh-Tuan Pham, *Member, IEEE*, Frédéric Vrins, *Graduate Student Member, IEEE*, and Michel Verleysen, *Senior Member, IEEE*

***Abstract*—Recently, some researchers have suggested Rényi's entropy in its general form as a blind source separation (BSS) objective function. This was motivated by two arguments: 1) Shannon's entropy, which is known to be a suitable criterion for BSS, is a particular case of Rényi's entropy, and 2) some practical advantages can be obtained by choosing another specific value for the Rényi exponent, yielding to, e.g., quadratic entropy. Unfortunately, by doing so, there is no longer guarantee that optimizing this generalized criterion would lead to recovering the original sources. In this paper, we show that Rényi's entropy in its exact form (i.e., out of any consideration about its practical estimation or computation) might lead to *not* recovering the sources, depending on the source densities and on Rényi's exponent value. This is illustrated on specific examples. We also compare our conclusions with previous works involving Rényi's entropies for blind deconvolution.**

***Index Terms*—Blind source separation (BSS), contrast function, independent component analysis, Rényi's entropy, Taylor expansion.**

I. Introduction

SINCE the early 1980s, the blind source separation (BSS) problem has revealed to be an important area of signal processing. It consists in recovering unknown source signals knowing only sensor recordings that are possibly noisy mixtures of them [16]. In its simplest form, the BSS model assumes that K sensors record linear and instantaneous mixtures $X_1(t),\ldots,X_K(t)$ of K-independent sources $S_1(t),\ldots,S_K(t)$; the mixing scheme is modeled by a real $K\times K$ invertible mixing matrix $\mathbf{A}$, so that, mathematically, we can write $\mathbf{X}=\mathbf{AS}$ where $\mathbf{X}=[X_1,\ldots,X_K]^T$ denotes the observed vector and $\mathbf{S}=[S_1,\ldots,S_K]^T$ is the source vector with independent components, ^{T} being the transpose symbol. The time index t has been omitted to simplify notation because the sources are assumed to be stationary. Under several assumptions, it is possible to recover the sources up to the amplitude and permutation indeterminacies; they correspond to the components Y_i of $\mathbf{Y}=\mathbf{BX}$ when $\mathbf{B}$ is obtained through the optimization of some criteria. One of them is the mutual information that can be written in terms of Shannon's entropies [5]. Recently, the use of a generalized form of Shannon's entropy, called Rényi's entropy, has been proposed to achieve the blind deconvolution and BSS problems [8], [13], [14], [23], [24]. In this last case, however, theoretical proofs ensuring that the sources will be recovered through the maximization of the related criteria were lacking. (The implicit conjecture that Rényi-entropy-based criterion is a contrast function has been corrected in a subsequent paper [9].) The analysis in [14] also suggests that Rényi-entropy-based criterion might not be appropriate for BSS in all circumstances. However, it is based on numerical calculation and is mostly restricted to Rényi's quadratic entropy and source distributions in the generalized Gaussian distribution family. The purpose of this paper is to provide a general theoretical analysis that yields a better understanding of the risk of using Rényi's entropy in its general form (that is, for all Rényi exponents not equal 1 or 0), at least in a totally blind scenario. Further, this work aims at answering why and when Rényi-entropy-based objective function may fail to yield the original sources. After a brief discussion in the next section regarding the use of entropies in BSS, the local maxima of the associated Rényi's criteria are analyzed in Section III through the first two derivatives of the criterion. Some detailed calculation on specific examples illustrates the above theoretical results in Section IV and shows that, depending on the cases, the use of Rényi's entropy as a BSS objective function may be risky, except when the Rényi exponent value is set equal to: 1) one and if at most one source is Gaussian, or 2) 0 if the sources have finite supports.

Manuscript received May 9, 2007; revised February 12, 2008. First published July 9, 2008; current version published September 17, 2008. The associate editor coordinating the review of this paper and approving it for publication was Prof. Philippe Loubaton. This article is an extended version of the paper "Is the general form of Rényi's entropy a contrast for source separation ?" presented at the 7th International Conference on Independent Component Analysis and Blind Signal Separation, London, U.K.

D.-T. Pham is with the Laboratoire Jean Kuntzmann, BP 53, 38041 Grenoble Cedex, France (e-mail: Dinh-Tuan.Pham@imag.fr).

F. Vrins is with the UCL Machine Learning Group, Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgium and also with the Financial Markets Department (Credit Trading Desk), ING South West Europe, 1000 Brussels, Belgium (e-mail: Frederic.Vrins@ing.be).

M. Verleysen is with the UCL Machine Learning Group, Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgium (e-mail: verleysen@dice.ucl.ac.be).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2008.928109

II. Generalized-Entropy-Based Criteria for BSS

The BSS problem can be addressed by two different methods: either the rows of $\mathbf{B}$ are estimated one by one (deflation BSS), or they are all estimated at once (simultaneous BSS).

- The deflation-based method tries to estimate, iteratively, an ith row $\mathbf{b}_i$ of $\mathbf{B}$ by

$$\mathbf{b}_i^\star=\arg\max_{\mathbf{b}_i}\bar{G}(\mathbf{b}_i\mathbf{X}) \tag{1}$$

subject to the constraint that $\mathrm{cov}(\mathbf{b}_i^\star\mathbf{X},\mathbf{b}_j^\star\mathbf{X})=0$ for $1\le j<i\le K$ where $\bar{G}$ is some suitable non-Gaussianity

measure. A simple choice for $\bar{G}$ is the absolute value of the kurtosis. Another possible choice is the exponential of the negentropy. For a random variable X of density p_X, its (Shannon) entropy is defined as $H(X) \doteq -E[\log p_X(X)]$ and its negentropy as $(1/2)\{\log[2\pi \mathrm{var}(X)]+1\} - H(X)$, which is the difference between the entropy of a Gaussian variable with variance $\mathrm{var}(X)$ and the entropy of X.

- The simultaneous BSS tries to globally estimate matrix $\mathbf{B}$. Adopting the mutual information-based criterion (which equals up to a constant $\sum_{k=1}^{K} H(\mathbf{b}_i\mathbf{X}) - \log|\det \mathbf{B}|$), one is led to the estimate

$$\mathbf{B}^{\star} = \arg\max_{\mathbf{B}} \left[\log|\det \mathbf{B}| - \sum_{k=1}^{K} H(\mathbf{b}_i\mathbf{X}) \right]. \qquad (2)$$

A generalized version of this criterion can be found in [20]: $\mathbf{B}$ is estimated as

$$\mathbf{B}^{*} = \arg\max_{\mathbf{B}} \left[\log|\det \mathbf{B}| - \sum_{i=1}^{K} \log Q(\mathbf{b}_i\mathbf{X}) \right] \qquad (3)$$

for some functional Q (possessing certain properties). Taking $Q(\cdot) = e^{H(\cdot)}$, the estimate (3) reduces to (2).

An "orthogonal version" of the above criterion can be derived: the search for $\mathbf{B}$ is restricted to the set $\mathbf{BCB}^T = \mathbf{I}_K$ where $\mathbf{C}$ denotes the covariance matrix of $\mathbf{X}$ and $\mathbf{I}_K$ the Kth-order identity matrix. This means that $\mathbf{W} = \mathbf{BA}$ is constrained to be in the orthogonal group of degree K, assuming without loss of generality that the independent sources have unit variance: $\mathbf{C} = \mathbf{A}E(\mathbf{SS}^T)\mathbf{A}^T = \mathbf{AA}^T$.

For the criterion to yield the sources (up to a permutation and a scaling), the set of its global maximum points must coincide with the set of monomial matrices

$$\mathcal{B} \doteq \{\mathbf{B} : \mathbf{BA} \text{ has a single nonzero element per row and column}\}.$$

If this is the case, the criterion is called a contrast function, in the sense of Comon [3].

In [20], Pham has proved that the criterion in (3) yields a *simultaneous* contrast function if $Q(\cdot)$ is strictly positive and is of class II superadditive in the sense of Huber [15], i.e., for any pair of random variables X and Y and two scalar numbers α and β

$$\begin{cases} Q(\alpha X + \beta) = |\alpha| Q(X), & \text{(class II)} \\ Q^2(X+Y) \geq Q^2(X) + Q^2(Y), & \text{(superadditive).} \end{cases}$$

Further, it can be shown that the *deflation* criterion in (1) is a contrast function for deflation scheme if $\bar{G}(X) = \mathrm{var}^{1/2}(X)/Q(X) = 1/Q[X/\mathrm{var}^{1/2}(X)]$ with Q a class II superadditive functional. In fact, $\bar{G}$ can be more general than that; it suffices that it is: 1) scale invariant (i.e., $\bar{G}(\alpha X) = \bar{G}(X)$ for any random variable X and any real number α) and 2) $\bar{G}(X+Y) \leq \max[\bar{G}(X), \bar{G}(Y)]$ for any pair of independent random variables X and Y, to ensure that this criterion is a contrast function [15].[1]

[1] $\bar{G}$, constructed as above, is clearly scale invariant and verifies $\bar{G}^2(X+Y) \leq [\mathrm{var}(X) + \mathrm{var}(Y)]/[Q^2(X) + Q^2(Y)] = \lambda \bar{G}^2(X) + (1-\lambda)\bar{G}^2(Y)$ where $\lambda = Q^2(X)/[Q^2(X) + Q^2(Y)]$, hence $\bar{G}^2(X+Y) \leq \max[\bar{G}^2(X), \bar{G}^2(Y)]$.

The class II superadditivity of the entropy power $e^{H(\cdot)}$ and of the Lebesgue measure of the support (or of the support convex hull, i.e., the range) can be proved using the entropy power inequality (EPI, [5]) and Brunn–Minkowski inequality (BMI, [5], [12]), respectively.[2] The log-measure of the support is also called the Hartley entropy [25]. Shannon's and Hartley's entropies are particular cases of Rényi's entropy, defined as [26]

$$H_r(X) \doteq \frac{1}{1-r} \log \int_{-\infty}^{\infty} p_X^r(x)dx, \qquad (r > 0). \qquad (4)$$

(In this paper, all densities are functions on the whole real line.) One can check that $\lim_{r\to 1} H_r(X) = H_1(X) = H(X)$ is the Shannon entropy and $\lim_{r\to 0} H_r(X) = H_0(X) = \log \mu(X)$, $\mu(X)$ being the Lebesgue measure of the support of X. Hence, from the above results, we can choose either $Q(\cdot) = e^{H(\cdot)}$ (if S_i has finite entropy) or $Q(\cdot) = \mu(\cdot)$ (if the sources have finite support measure).

Rényi's entropy is decreasing and continuous in r [18]. The major properties of Rényi's entropy are the same as Shannon's. In particular, if α and β are two scalar numbers, $\mathbf{M}$ is a (deterministic) square matrix and $\mathbf{U}$ is a (deterministic) vector of the same size as that of the random vector $\mathbf{X}$

$$\begin{cases} H_r(\alpha X + \beta) = H_r(X) + \log|\alpha| \\ H_r(\mathbf{MX} + \mathbf{U}) = H_r(\mathbf{X}) + \log|\det \mathbf{M}|. \end{cases}$$

Consequently, $e^{H_r(\cdot)}$ is a class II functional.

It was first our hope to find a generalized form of EPI and BMI that would ensure that $e^{H_r(\cdot)}$ with arbitrary $r > 0$ would also be superadditive; Indeed, if it were the case, one could take $Q(\cdot) = e^{H_r(\cdot)}$ and obtain a contrast function, as previously explained. Unfortunately, no such result could be found (in fact, this functional is not superadditive for $r \notin \{0, 1\}$ as it will be seen below). Therefore, instead of trying to prove that the Rényi-entropy-based criterion is a contrast function, we will check if it satisfies some necessary conditions to be so. If it does not, then it is not a contrast function (and consequently from [20], the functional $Q(\cdot) = e^{H_r(\cdot)}$ is not superadditive). In the deflation case, a necessary condition that the criterion in (1) is a contrast function is that $\bar{G}(\mathbf{bX})$ attains a local maximum at some ith row of $\mathbf{A}^{-1}$, or equivalently, $Q(S_i)$ for some i is local minimum, under the constraint $\mathbf{bCb}^T = 1$, of $Q(\mathbf{bX})$. In the simultaneous approach, a necessary condition for the criterion in (3) to be a contrast function is that this criterion attains a local maximum at $\mathbf{B} = \mathbf{A}^{-1}$. In Section III, we will investigate the local maxima of the Rényi-entropy-based criterion, via a Taylor expansion of $H_r(\cdot)$ up to the second order.

III. On Local Maxima of the Criterion

In this section, we consider the BSS criteria involved in (1) and (3) with $\bar{G}(\cdot) = \mathrm{var}^{1/2}(\cdot)/e^{H_r(\cdot)}$ and $Q(\cdot) = e^{Hr(\cdot)}$. The first-order Taylor expansion of the entropy will be computed in Section III-A. This result will be helpful when analyzing the stationary points of the criteria. In Section III-B, these stationary points will be further characterized using a second-order expansion around them.

[2] Interestingly, some relationships between the above inequalities have been pointed out [4].

A. First-Order Analysis

We start by computing the first-order approximation of the Rényi's entropy of $Y+\mathbf{hZ}$, a random variable Y contaminated by a small random variable of the form $\mathbf{hZ}$, where $\mathbf{Z}$ is a random column vector and $\mathbf{h}$ is a small row vector (in terms of its Euclidean norm $\|\mathbf{h}\|$).[3] Let p_Y be the density of Y. It will be shown in the Appendix that the Rényi's entropy of $Y+\mathbf{hZ}$ admits the first-order Taylor expansion[4]

$$H_r(Y+\mathbf{hZ}) = H_r(Y)+\mathbf{h}E[\psi_{Y,r}(Y)\mathbf{Z}]+o(\|\mathbf{h}\|), \qquad \mathbf{h}\to 0 \tag{5}$$

where

$$\psi_{Y,r} \doteq -\frac{rp_Y^{r-2}p_Y'}{\int p_Y^r(y)dy} = -\frac{(p_Y^r)'}{\int p_Y^r(y)dy}\frac{1}{p_Y} \tag{6}$$

which we call the rth score function. Note that for $r=1$, the rth score function is simply the score function and (5) reduces to the well-known first-order expansion of the Shannon's entropy provided in [22]. It is of interest to note that

$$E[\psi_{Y,r}(Y)] = 0 \quad E[\psi_{Y,r}(Y)Y] = 1. \tag{7}$$

The first equality follows directly from the definition (6) of $\psi_{Y,r}$ and the second equality is obtained through integration by parts.

The above result will be useful for analyzing the stationary points of the Rényi-entropy-based criteria for BSS. We consider below the deflation and simultaneous approaches separately.

1) Deflation Approach: Put $\mathbf{w}=\mathbf{bA}$ so that $H_r(\mathbf{bX}) = H_r(\mathbf{wS})$. For a small increment $\boldsymbol{\delta} \doteq [\delta_1,\ldots,\delta_K]$ of $\mathbf{w}$, one gets from (5)

$$-H_r(\mathbf{wS}+\boldsymbol{\delta}\mathbf{S}) = -H_r(\mathbf{wS}) - \sum_{k=1}^{K}\delta_k E[\psi_{\mathbf{wS},r}(\mathbf{wS})S_k]+o(\|\boldsymbol{\delta}\|). \tag{8}$$

Further, the constraint $\mathbf{bCb}^T=1$ translates to $\|\mathbf{w}\|^2=1$, that is, the incremented vector must satisfy $\|\mathbf{w}+\boldsymbol{\delta}\|^2=1$. This yields $\mathbf{w}\boldsymbol{\delta}\mathbf{w}^T = -(1/2)\|\boldsymbol{\delta}\|^2$. Thus, if $\mathbf{w}=\pm\mathbf{e}_j$, where $\mathbf{e}_j$ denotes the jth row $\mathbf{I}_K$, then $\delta_j = o(\|\boldsymbol{\delta}\|)$, and from (8) and noting that $E[\psi_{\pm S_j}(\pm S_j)S_k] = \pm 1$ for $k=j$ and 0 otherwise [by (7) and the independence of the sources], one gets

$$-H_r(\pm S_j+\boldsymbol{\delta}\mathbf{S}) = -H_r(\pm S_j)+o(\|\boldsymbol{\delta}\|). \tag{9}$$

Thus, $\exp[-H_r(\mathbf{bX})]$ admits, on the set $\{\mathbf{b}:\mathbf{bCb}^T=1\}$, a stationary point at $\pm\mathbf{e}_j\mathbf{A}^{-1} = \pm(\mathbf{A}^{-1})_{j\cdot}$, $\forall\ j$, where $(\mathbf{A}^{-1})_{j\cdot}$ denotes the jth row of $\mathbf{A}^{-1}$. Because $\bar{G}_r(\mathbf{bX}) = \mathrm{var}^{1/2}(\mathbf{bX})/\exp[H_r(\mathbf{bX})]$ is scale invariant, as a function of $\mathbf{b}$ without constraint, it admits a stationary point at any multiple of a row of $\mathbf{A}^{-1}$.

[3]In this section, no assumption about the possible dependence between Y and $\mathbf{Z}$ is required.

[4]In this paper, we adopt the following convention regarding the expectations to simplify the notations: the expectation of a function of variables is always taken according to each of the variables involved in the function.

2) Simultaneous Approach: The simultaneous criterion associated to $H_r(\cdot)$ is

$$C_r(\mathbf{B}) \doteq \log|\det\mathbf{B}| - \sum_{i=1}^{K} H_r(\mathbf{b}_i\mathbf{X}). \tag{10}$$

It follows that for a small matrix $\mathcal{E}$

$$C_r(\mathbf{B}+\mathcal{E}\mathbf{B}) = \log|\det(\mathbf{B}+\mathcal{E}\mathbf{B})| - \sum_{i=1}^{K} H_r[(\mathbf{B}+\mathcal{E}\mathbf{B})_{i\cdot}\mathbf{X}]. \tag{11}$$

Recall that $\log|\det(\mathbf{B}+\mathcal{E}\mathbf{B})| = \log|\det\mathbf{B}|+\mathrm{tr}(\mathcal{E}) - (1/2)\mathrm{tr}(\mathcal{E}^2)+o(\|\mathcal{E}\|^2)$, where $\mathrm{tr}(\cdot)$ denotes the *trace* operator. On the other hand, let $\mathcal{E}_{ij}$ be the general element of $\mathcal{E}$

$$(\mathbf{B}+\mathcal{E}\mathbf{B})_{i\cdot}\mathbf{X} = \left(\mathbf{b}_i+\sum_{j=1}^{K}\mathcal{E}_{ij}\mathbf{b}_j\right)\mathbf{X} = (\mathbf{w}_i+\boldsymbol{\delta})\mathbf{S}$$

where we have put $\mathbf{w}_i=\mathbf{b}_i\mathbf{A}$ and $\boldsymbol{\delta}=\sum_{j=1}^{K}\mathcal{E}_{ij}\mathbf{w}_j$. Thus, applying (8), one gets

$$-H_r[(\mathbf{B}+\mathcal{E}\mathbf{B})_{i\cdot}\mathbf{X}] = -H_r(\mathbf{w}_i\mathbf{S}) - \sum_{k=1}^{K}\delta_k E[\psi_{\mathbf{w}_i\mathbf{S},r}(S_k)]+o(\|\boldsymbol{\delta}\|).$$

In the sequel, we consider the particular case where $\mathbf{BA}$ is a matrix with a single nonzero entry per row and column, that is, $\mathbf{BA}\in\mathcal{B}$, or equivalently, of the form $\mathbf{PD}$ where $\mathbf{P}$ is a permutation matrix and $\mathbf{D}$ is a diagonal matrix. Then, $\mathbf{w}_i$ is of the form $d_{\pi_i}\mathbf{e}_{\pi_i}$ for some permutation $\pi_1,\ldots,\pi_K$ of $1,\ldots,K$ and some real numbers $d_1,\ldots,d_K$; this yields $Y_i \doteq \mathbf{w}_i\mathbf{S} = d_{\pi_i}S_{\pi_i}$ and $\delta_{\pi_k}=\mathcal{E}_{ik}d_{\pi_k}$. Then, using the independence of the sources and (7), the above expansion becomes

$$-H_r[(\mathbf{B}+\mathcal{E}\mathbf{B})_{i\cdot}\mathbf{X}] = -H_r(Y_i) - \mathcal{E}_{ii} + o\left[\left(\sum_{k=1}^{K}\mathcal{E}_{ik}^2\right)^{1/2}\right].$$

Combining the above results, one gets finally that $C_r((\mathbf{I}_K+\mathcal{E})\mathbf{PDA}^{-1}) = C_r(\mathbf{PDA}^{-1})+o(\|\mathcal{E}\|)$. This shows that any matrix of the form $\mathbf{PDA}^{-1}$ (that is, any point of $\mathcal{B}$) is a stationary point of the criterion C_r. For the orthogonal version of this criterion, $\mathbf{B}$ is constrained to satisfy $\mathbf{BCB}^T=\mathbf{I}_K$. It can still be of the form $\mathbf{PDA}^{-1}$ but with the diagonal matrix $\mathbf{D}$ having diagonal elements ± 1. The matrix $\mathcal{E}$ must be such that $\mathbf{B}+\mathcal{E}\mathbf{B}$ also satisfy the constraint, which is equivalent to $\mathbf{I}_K+\mathcal{E}$ being an orthogonal matrix. Then, the above expansion for $C_r(\mathbf{B}+\mathcal{E}\mathbf{B})$ remains valid under these constraints, showing that $\mathbf{PDA}^{-1}$ is still a stationary point of the orthogonal version of the criterion.

B. Second-Order Analysis

It is shown in the above section that $\bar{G}_r(\mathbf{bX})$ admits a stationary point at any multiple of a row of $\mathbf{A}^{-1}$ and $C_r(\mathbf{B})$ admits a stationary point at any point in $\mathcal{B}$. At this step, however, we are not able to further characterize these points: Do they correspond to a local maximum or a minimum or saddle points? To

answer this question, we have to extend the Taylor expansion of the criterion up to the second order around these specific points.

As in Section III-A, we start by expanding $H_r(Y+\mathbf{hZ})$, but unlike in that section, we will limit ourselves to the case where $\mathbf{Z}$ is independent of Y, which is enough for our purpose and allows a much simpler calculation. It is shown in the Appendix that

$$H_r(Y+\mathbf{hZ}) = H_r(Y) + \frac{1}{2}J_r(Y)\mathrm{var}(\mathbf{hZ}) + o(\|\mathbf{h}\|^2), \quad \mathbf{h}\to 0 \quad (12)$$

where

$$J_r(Y) \doteq \frac{r}{1-r}\frac{\int p_Y^{r-1}(y)p_Y'(y)dy}{\int p_Y^r(y)dy} = r\frac{\int p_Y^{r-2}(y)p_Y^2(y)dy}{\int p_Y^r(y)dy}. \quad (13)$$

In (13), the last equality comes from the integration by parts. We call $J_r(Y)$ the rth Fisher information of Y. Observe that $J_r(Y)$ can also be rewritten as $E[\psi_Y(Y)\psi_{r,Y}(Y)]$; however, the right-hand side of (13) makes it easier to see that $J_r(Y)>0$. For $r=1$, $J_1(Y)$ is no other than Fisher's information $J(Y)$ of Y [5].

We use the above result to analyze the second-order properties of the Rényi-entropy-based criteria for BSS. As we did when analyzing their stationary points, we will consider the deflation and the simultaneous approaches separately.

1) Deflation Approach: We will extend the entropy expansion in (8) up to the second order around $\mathbf{w}_j=\mathbf{e}_j$. From (12), one gets

$$-H_r(\pm S_j+[\delta_1,\ldots,\delta_K]\mathbf{S}) = -H_r[(\delta_j\pm 1)S_j] - \frac{1}{2}\sum_{1\le k\ne j\le K}\delta_k^2 J_r(\pm S_j) + o\left(\sum_{1\le k\ne j\le K}\delta_k^2\right).$$

Note that $J_r(S_j)=J_r(-S_j)$ and $H_r[(\delta_j\pm 1)S_j]=H_r(\pm S_j)+\log|1\pm\delta_j|$, and further, the unit norm constraint yields $(1\pm\delta_j)^2 = 1-\sum_{1\le k\ne j\le K}\delta_k^2 \le 1-\sum_{k=1}^K\delta_k^2$. Therefore

$$-H_r(\pm S_j+\boldsymbol{\delta}\mathbf{S}) = -H_r(\pm S_j) - \frac{1}{2}\times\sum_{1\le k\ne j\le K}\delta_k^2[J_r(S_j)-1] + o(\|\boldsymbol{\delta}\|^2).$$

The above result shows that a necessary condition for the function $-H_r(\mathbf{wS})$ to admit a local maximum at $\pm\mathbf{e}_j$ over the set $\{\mathbf{w}\in\mathbb{R}^K : \|\mathbf{w}\|=1\}$ is that $J_r(S_j)\ge 1$ and a sufficient condition is that this inequality is strict. Because the sources have been assumed to have unit variance, one can write these conditions as $J_r(S_j)\mathrm{var}(S_j)\ge 1$ and $J_r(S_j)\mathrm{var}(S_j)>1$, which are then independent of the source variance.

From the above results, we conclude the following.

Theorem 1: The criterion $\bar{G}_r(\mathbf{bX}) = \mathrm{var}^{1/2}(\mathbf{bX})/\exp[H_r(\mathbf{bX})] = \exp\{-H_r[\mathbf{bX}/\mathrm{var}^{1/2}(\mathbf{bX})]\}$ is not a (deflation) contrast function if $J_r(S_i)\mathrm{var}(S_i)<1$ where $i=\arg\max_{k\in\{1,\ldots,K\}} -H_r\left[S_k/\mathrm{var}^{1/2}(S_k)\right]$.

2) Note: The above result applies to the first step of the deflation approach where the source with the highest $\bar{G}_r$ has to be extracted. If one applies successively the deflation approach to extract all sources, then the method will fail if $J_r(S_i)\mathrm{var}(S_i)<1$ for only a single index i (in fact, it fails to extract the source of this index).

3) Simultaneous Approach: We now derive a necessary and sufficient condition for the criterion $C_r(\mathbf{B})$ to attain a local maximum at point $\mathbf{B}=\mathbf{PDA}^{-1}$, where $\mathbf{P}$ is a permutation matrix and $\mathbf{D}$ is a diagonal matrix.

Let Y_i denote the ith component of $\mathbf{BX}$, then the ith component of $(\mathbf{B}+\mathcal{E}\mathbf{B})\mathbf{X}$ can be written as $Y_i+\sum_{j=1}^K\mathcal{E}_{ij}Y_j$, where $\mathcal{E}_{ij}$ is the general element of $\mathcal{E}$. For $\mathbf{B}$ of the above form, $Y_i=d_{\pi_i}S_{\pi_i}$ for some permutation $\pi_1,\ldots,\pi_K$ of $1,\ldots,K$ and some nonzero real numbers $d_1,\ldots,d_K$. Thus, the Y_k are independent and the expansion (12) yields

$$H_r\left(Y_i+\sum_{j=1}^K\mathcal{E}_{ij}Y_j\right) = H_r[(1+\mathcal{E}_{ii})Y_i] + \frac{1}{2}\sum_{1\le j\ne i\le K}\mathcal{E}_{ij}^2\mathrm{var}(Y_j)J_r(Y_i) + o\left(\sum_{j=1}^K\mathcal{E}_{ij}^2\right).$$

Note that $H_r[(1+\mathcal{E}_{ii})Y_i] = H_r(Y_i)+\log(1+\mathcal{E}_{ii}) = H_r(Y_i)+\mathcal{E}_{ij}-(1/2)\mathcal{E}_{ij}^2+o\left(\mathcal{E}_{ij}^2\right)$. It follows from the above results and the expansion of $\log\left|\det(\mathbf{B}+\mathcal{E}\mathbf{B})\right|$ that

$$C_r(\mathbf{B}+\mathcal{E}\mathbf{B}) = C_r(\mathbf{B}) - \frac{1}{2}\sum_{1\le i\ne j\le K}\sum\left[\mathcal{E}_{ij}^2J_r(Y_i)\mathrm{var}(Y_j)+\mathcal{E}_{ij}\mathcal{E}_{ji}\right] + o(\|\mathcal{E}\|^2).$$

The second sum in the above right-hand side is a quadratic form associated with the symmetric block diagonal matrix, with 2×2 blocks

$$\mathbf{J}^{i,j} \doteq \begin{bmatrix} J_r(Y_i)\mathrm{var}(Y_j) & 1 \\ 1 & J_r(Y_j)\mathrm{var}(Y_i)\end{bmatrix} \quad (14)$$

that is

$$\sum_{1\le i\ne j\le K}\sum\left[\mathcal{E}_{ij}^2J_r(Y_i)\mathrm{var}(Y_j)+\mathcal{E}_{ij}\mathcal{E}_{ji}\right] = \sum_{1\le i<j\le K}\sum[\,\mathcal{E}_{ij}\quad\mathcal{E}_{ji}\,]\,\mathbf{J}_r^{i,j}\,[\,\mathcal{E}_{ij}\quad\mathcal{E}_{ji}\,]^T.$$

Thus, in order for the criterion $C_r(\mathbf{B})$ to reach a local maximum at point $\mathbf{B}=\mathbf{PDA}^{-1}$ (i.e., $\mathbf{B}\in\mathcal{B}$), it is necessary that the $\mathbf{J}^{i,j}$ matrices in (14) are positive semidefinite and it is sufficient that they are positive definite. However, $\mathbf{J}^{i,j}$ is positive definite if and only if all of its eigenvalues are strictly positive. This is equivalent to $\mathrm{tr}(\mathbf{J}^{i,j})>0$ and $\det(\mathbf{J}^{i,j})>0$. Because $J_r(Y)$ is a positive quantity, the necessary condition for the criterion $C_r(\mathbf{B})$ to attain a local maximum at $\mathbf{PDA}^{-1}$ is $J_r(Y_i)\mathrm{var}(Y_i)J_r(Y_j)\mathrm{var}(Y_j)\ge 1, \forall i\ne j$. The sufficient condition is that the above inequality is strict.

In the case where the sources have the same distribution as that of some random variable S, the above necessary condition reduces to $J_r(S)\mathrm{var}(S)\ge 1$. The sufficient condition is $J_r(S)\mathrm{var}(S)>1$.

Hence, the following theorem has been proved.

Theorem 2: The criterion $C_r(\mathbf{B})$ is not a contrast function if there exists a pair of indices $i, j \in \{1, \ldots, K\}$ such that the inequality $J_r(S_i)\mathrm{var}(S_i)J_r(S_j)\mathrm{var}(S_j) < 1$ holds true.

As a corollary, a simplified version of this theorem may be found by assuming that all the sources share the same density.

Corollary 1: The criterion $C_r(\mathbf{B})$ is not a contrast function if the sources share the same distribution as that of some random variable S and $J_r(S)\mathrm{var}(S) < 1$.

Theorems 1 and 2 give sufficient conditions ensuring that the above deflation and simultaneous criteria are *not* contrast functions: maximizing them will *not* lead to obtaining the sources if these conditions hold. It is shown in the next section that these conditions can hold true for densities close to (but different from) the Gaussian density and for any values of Rényi's exponent r and for the specific densities and values of this exponent.

IV. Detailed Results for Specific Source Densities

Consider the case where the sources admit a common density belonging to the family of the generalized Laplace (or Gaussian) distribution

$$p_S(s) = C \exp\left(-\frac{1}{a}\left|\frac{s}{\lambda}\right|^a\right) \tag{15}$$

where a is a positive parameter, λ is a positive scale parameter, and C is the normalizing constant. Then, the rth score function of the random variable S with density p_S reduces to

$$\psi_{S,r}(y) = \frac{r\,\mathrm{sign}(y)|y|^{a-1}\lambda^{-a}}{C\exp\left[-\dfrac{1-r}{a}\left|\dfrac{y}{\lambda}\right|^a\right]\int \exp\left(-\dfrac{r}{a}\left|\dfrac{u}{\lambda}\right|^a\right)du} \tag{16}$$

for $r > 0$. In particular, $\psi_{S,1}(y) = \mathrm{sign}(y)|y|^{a-1}\lambda^{-a}$. Further

$$\begin{aligned} J_r(S) &= \frac{r\int |s|^{2a-2}\lambda^{-2a}\exp\left(-\dfrac{r}{a}\left|\dfrac{s}{\lambda}\right|^a\right)ds}{\int \exp\left(-\dfrac{r}{a}\left|\dfrac{s}{\lambda}\right|^a\right)ds} \\ &= \frac{r\int |u|^{2a-2}\exp\left(\dfrac{-r|u|^a}{a}\right)du}{\lambda^2\int \exp\left(\dfrac{-r|u|^a}{a}\right)du} \\ &= \frac{r^{2/a-1}}{\lambda^2}\frac{\int |z|^{2a-2}\exp\left(\dfrac{-|z|^a}{a}\right)du}{\int \exp\left(\dfrac{-|z|^a}{a}\right)dz} \\ &= \frac{r^{2/a-1}}{\lambda^2}E|Z|^{2a-2} \end{aligned} \tag{17}$$

where $Z = S/\lambda$ is a random variable with density $\exp(-|z|^a/a)/\int \exp(-|u|^a/a)du$. Because $\mathrm{var}(S) = \lambda^2 E(Z^2)$, one has

$$J_r(S)\mathrm{var}(S) = r^{2/a-1}E(|Z|^{2a-2})E(Z^2), \qquad (r > 0) \tag{18}$$

which is independent of the scale parameter λ as it should be. In particular, for $a = 2$, which corresponds to S and Z being Gaussian (with $E(Z^2) = 1$), one has $J_r(S)\mathrm{var}(S) = 1$, $\forall\ r > 0$. This is, of course, expected because one cannot separate Gaussian sources.

From the above result, $J_r(S)\mathrm{var}(S) = r^{2/a-1}J(S)\mathrm{var}(S)$, where $J(S) = J_1(S)$ is the Fisher information of S, and $J(S)\mathrm{var}(S) = E(|Z|^{2a-2})E(Z^2)$, which we also denote by $g(a)$ to emphasize its dependence on a. We know by using the Cramér–Rao inequality that $J(S)\mathrm{var}(S) \geq 1$ with equality if and only if S is Gaussian, that is, $a = 2$, hence g admits a global minimum equal to 1 at $a = 2$. Thus, $g(a) > 1$ for all $a \neq 2$. It follows that $J_r(S)\mathrm{var}(S) < 1$ if and only if

$$r < g(a)^{1/(1-2/a)} < 1 \quad \text{in the case } a < 2 \tag{19}$$

$$r > g(a)^{1/(1-2/a)} > 1 \quad \text{in the case } a > 2. \tag{20}$$

These results are summarized in Theorem 3.

Theorem 3: For a source density of the form $p_S(s) = C\exp(-|s/\lambda|^a/a)$, if $a < 2$, the criteria are not contrast functions for $r < g(a)^{a/(a-2)}$, and if $a > 2$, the criteria are not contrast functions for $r > g(a)^{a/(a-2)}$. Furthermore, for *any given* r distinct from 1 and 0, there exists a source density of the form $C\exp(-|s/\lambda|^a/a)$ for some a for which the contrast properties are not met by the criteria.

Proof: The first part of the theorem results from the above development. To prove that a generalized Laplacian density-based counterexample can be found, whatever r is, consider the function $a \mapsto r^{2/a-1}g(a) = J_r(S)\mathrm{var}(S)$. It takes the value 1 at 2 and its logarithmic derivative is $-2a^{-2}\log r + g'(a)/g(a)$, which takes the value $-(1/2)\log r$ at 2 (because g is minimum at 2, hence $g'(2) = 0$). Thus, for $r < 1$, this function is increasing in a neighborhood of 2, hence there exists an $a < 2$ for which $J_r(S)\mathrm{var}(S) < 1$. Similarly, for $r > 1$, this function is decreasing in a neighborhood of 2, hence there exists an $a > 2$ for which $J_r(S)\mathrm{var}(S) < 1$. This concludes the proof of the theorem. □

The function g can be computed explicitly. From the definition of the gamma function $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$, one has

$$\begin{aligned} E(|Z|^\beta) &= \frac{\int_0^\infty \exp\left(\dfrac{-z^a}{a}\right)z^\beta dz}{\int_0^\infty \exp\left(\dfrac{-z^a}{a}\right)dz} \\ &= \frac{\int_0^\infty \exp(-t)(at)^{(\beta+1)/a-1}dt}{\int_0^\infty \exp(-t)(at)^{1/a-1}dt} \\ &= a^{\beta/a}\frac{\Gamma\left[\dfrac{(\beta+1)}{a}\right]}{\Gamma\left(\dfrac{1}{a}\right)} \end{aligned} \tag{21}$$

and, therefore

$$g(a) = E(|Z|^{2a-2})E(Z^2) = a^2\frac{\Gamma\left(\dfrac{2-1}{a}\right)\Gamma\left(\dfrac{3}{a}\right)}{\Gamma\left(\dfrac{1}{a}\right)^2}. \tag{22}$$

Applying the above theorem to bilateral exponential sources, which correspond to $a = 1$, one has $g(a) = 2$, hence the criteria are not contrast functions for $r < 1/2$. For $a = 4$,

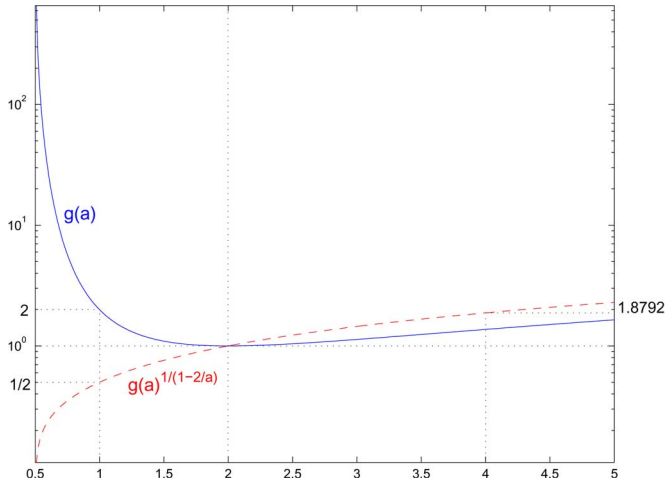

Fig. 1. Functions $g(a)$ (solid) and $g(a)^{1/(1-2/a)}$ (dashed); the specific values at $a = 4$ and $a = 1$ are pointed out.

$g(a)^{1/(1-2/a)} = 1.8792$, and thus, the criteria are not contrast functions for $r > 1.8792$, in particular, for $r = 2$ (which corresponds to the quadratic Rényi entropy). The functions $g(a)$ and $g(a)^{1/(1-2/a)}$ are illustrated in Fig. 1. The plot of $g(a)$ confirms the theoretical result that g admits a global minimum equal to 1 at $a = 2$. The plot of $g(a)^{1/(1-2/a)}$ indicates the range of values of r for which the criteria are not contrast functions.

As a numerical illustration, we consider the case of two unit-variance sources sharing the same density p_S given in (15). To fulfill the $||\mathbf{w}|| = 1$ constraint, we set $\mathbf{w} = \mathbf{w}_\theta \doteq [\sin\theta, \cos\theta]$. We compute numerically the Rényi's entropy $H_r(\mathbf{w}_\theta \mathbf{S})$ (solid curves) as a function of the transfer angle θ for the two values of a and plot the results on Fig. 2(a) and (b) (see below for the calculation method).

Remark 1 (Some Details Regarding the Numerical Calculation): $H_r\left(\mathbf{w}_\theta [S_1, S_2]^T\right)$ is computed numerically as shown in the equation at the bottom of the page, where the $\sum_\Delta$ symbol denotes the Riemannian approximation of the exact integral (the step Δ is taken equal to 0.002 and the grid size is chosen large enough to ensure that the integration error is limited to $\max(|1 - \sum_\Delta p_{\sin\theta S}|, |1 - \sum_\Delta p_{\cos\theta S}|) < \tau$, $\tau = \Delta/2$. Similarly, the variance deviation error is also maintained below $\max\left(\left|1 - \left(\sum_\Delta s^2 p_{\sin\theta S}(s)\right)\right|, \left|1 - \left(\sum_\Delta s^2 p_{\cos\theta S}(s)\right)\right|\right) < \tau$). The exact theoretical expressions of $p_{\sin\theta S}$ and $p_{\cos\theta S}$ have been used and the convolution operation (denoted by the "$\star$" symbol) is performed via the MatLab*conv* command.

Remark 2: Fig. 2(a) and (b) *seems* to indicate that even if the kind of the extremum points change with r (maximum or minimum), their location does not change when r varies. This is not the case in general. It has been shown that a stationary point always exists when $\theta = \pi/2$, whatever r is. Simple calculation yields, noting $Y_\theta = \mathbf{w}_\theta \mathbf{S}$

$$Y_{\theta+\Delta\theta} = Y_\theta + [\Delta\theta\cos\theta, -\Delta\theta\sin\theta]\mathbf{S} + o(\Delta\theta).$$

From (8), this leads to

$$-H_r(Y_{\theta+\Delta\theta}) = -H_r(Y_\theta) - \Delta\theta\cos\theta E[\psi_{Y_\theta,r} S_1] + \Delta\theta\sin\theta E[\psi_{Y_\theta,r} S_2] + o(\Delta\theta).$$

Consequently, $H_r(Y_\theta)$ admits a stationary point at $\theta^\star$ if

$$\tan\theta^\star = \frac{E[\Psi_{Y_{\theta^\star},r} S_1]}{E[\Psi_{Y_{\theta^\star},r}(Y_{\theta^\star}) S_2]} = \frac{\int \left(p^r_{Y_{\theta^\star}}\right)'(y) E[S_1 | Y_{\theta^\star} = y] dy}{\int \left(p^r_{Y_{\theta^\star}}\right)'(y) E[S_2 | Y_{\theta^\star} = y] dy}.$$

In general, $\theta^\star$ depends on r.

As the last example, consider the case where the sources admit a common density $p_S(s) = 1 - |s|$ if $|s| \le 1$, $= 0$, otherwise. Then

$$\mathrm{var}(S) = 2\int_0^2 (1-s)s^2 ds = \frac{1}{6} \tag{23}$$

and[5]

$$J_r(S) = r\frac{\int_0^1 (1-s)^{r-2} ds}{\int_0^1 (1-s)^r ds} = r\frac{\int_0^1 u^{r-2} du}{\int_0^1 u^r du} = \begin{cases} \dfrac{r(r+1)}{(r-1)}, & \text{if } r > 1 \\ \infty, & \text{if } r \le 1. \end{cases}$$

Thus, $J_r(S)\mathrm{var}(S) < 1$ if and only if $r > 1$ and $r(r+1)/[6(r-1)] < 1$. However, for $r \ge 1$, the last inequality is equivalent to $0 > r(r+1) - 6(r-1) = (r-2)(r-3)$. Therefore, $J_r(S)\mathrm{var}(S) < 1$ if and only if $2 < r < 3$. We conclude that for the case of two triangular sources, the criteria are not contrast functions if $2 < r < 3$. This is shown in Fig. 3 in the simultaneous case. Note that even if we know that the criterion is not a contrast function for r in this interval because a necessary condition is not met, there is no guarantee that it would be a contrast function for values of r outside this interval: the global maximum may be reached at some mixing point, even if we know that the nonmixing points maximize the criterion locally (see $r = 5$ in Fig. 3).

[5]The triangular density is piecewise differentiable and continuous. Even if its derivative has the jumps at some isolated points, (25) is still valid if such points are excluded and (27) still holds if the second derivative and convergence is understood in the sense of the distributions (or generalized functions). One can actually prove that (12) is still valid for $r > 1$, provided that the r-Fisher information is defined by the second right-hand side of (13).

$$\begin{cases} (1-r)^{-1}\log\sum_\Delta [p_{\sin\theta S} \star p_{\cos\theta S}]^r & \text{if } \theta \notin \left\{\dfrac{k\pi}{2}, k \in \mathbb{Z}\right\} \\ (1-r)^{-1}\log\sum_\Delta p_S^r & \text{otherwise} \end{cases}$$

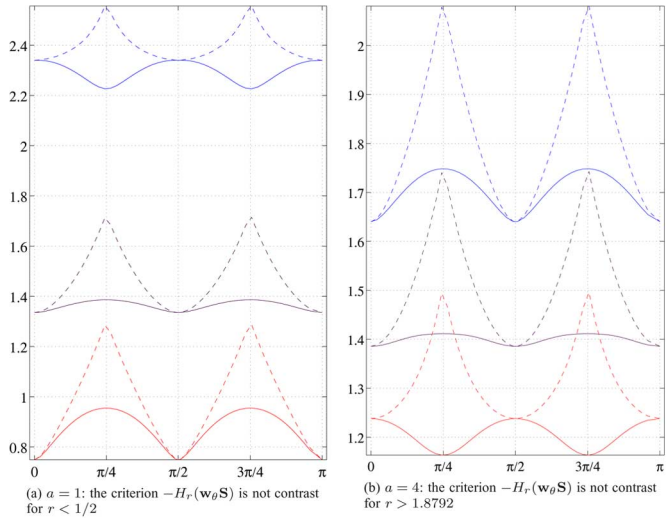

Fig. 2. Evolution of $H_r(\mathbf{w}_\theta \mathbf{S})$ (solid lines, criterion to be minimized) and $H_r(Y_\theta / \max(\sin(\theta), \cos(\theta)))$ (dashed lines, see Section V) where $p_{S_1} = p_{S_2} = p_S$ is given by (15); the curves are shown for $r = 0.2$ (top), $r = 1$ (middle) and $r = 5$ (bottom). (Recall that H_r is decreasing with r.) The contrast property of $-H_r(Y_\theta)$ (criterion to be maximized) depends on r and a; on the contrary, because of the addition of an extra term, $H_r[Y_\theta / \max(|\sin\theta|, |\cos\theta|)] = H_r(Y_\theta) - \log\max(|\sin\theta|, |\cos\theta|)$ is always minimized at 0, $\pi/2$, π, whatever r and a are.

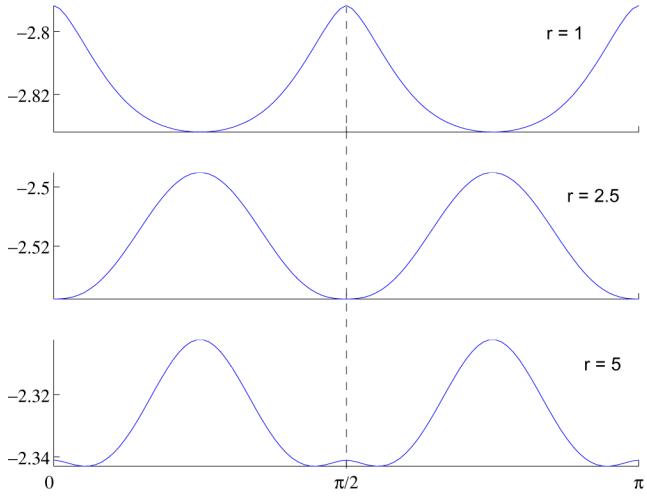

Fig. 3. Evolution of estimated Rényi's criterion $-H_r(Y_\theta) - H_r(Y_{\pi/2-\theta})$ (to be maximized) as a function of the transfer angle θ where the two sources share the same triangular density. The criterion with $r = 2.5$ and $r = 5$ is not a contrast function because there is no local maximum at the nonmixing points and because the local maximum is not global, respectively.

V. Discussion and Comparison With Existing Results

The general form of the Rényi's entropy has been proposed for BSS [8], [13], [14], [24]. In the linear instantaneous case, however, only the entropies with exponent $r = 0$ [7], [19], [29], [30] or $r = 1$ [3], [6], [21] have been proved to yield a contrast function. Close relationship with Shannon's and Hartley's entropies was the motivations for the use of the Rényi's entropy with $r \notin \{0, 1\}$. The quadratic case ($r = 2$) was also motivated by computational convenience and some other reasons given in [14]. In the above referenced papers, simulations suggest that maximizing Rényi-entropy-based criteria may be suitable for BSS. However, they involve only specific values of r (mostly $r = 2$) and specific source densities, whereas our approach relies on a wider theoretical analysis. Our results complete the above empirical studies and show that using the Rényi's entropy for source separation in a *blind* context runs the risk of finding the mixture, because the Rényi's entropy may *not* lead to a contrast function in general.

Rényi's entropy was also proposed for the blind deconvolution problem [2], [10], [11]. The justification provided in the referenced papers relies on the following inequality. For independent random variables $S_1, \ldots, S_K$ with the *same distribution* as some random variable S

$$H_r\left(\sum_{i=1}^{K} w_i S_i\right) \geq H_r(S) + \log \max_{i=1,\ldots,K} |w_i| \qquad (24)$$

with equality if and only if all but one of the w_i are zero. The requirement that $S_1, \ldots, S_K$ have the same distribution is natural in the deconvolution context because the S_i represent the (stationary) source at different sample points. The variable $Y \doteq \sum_{i=1}^{K} w_i S_i$ then represents the output of the deconvolution filter. The above inequality thus shows that the source can be recovered (deconvolved) up to a delay by minimizing $H_r(Y)$ under the $\max_{i=1,\ldots,K} |w_i| = 1$ constraint.

One may try to apply the above argument to build a *deflation* contrast function for BSS in the case where the sources have the same distribution, which would apparently contradict our result. However, the above argument relies on changing the constraint $\sum_{i=1}^{K} w_i^2 = 1$ into $\max_{i=1,\ldots,K} |w_i| = 1$, which is, strictly speaking, not really a *true constraint* because it depends on the unknown mixing matrix $\mathbf{A}$. Note that the constraint $\sum_{i=1}^{K} w_i^2 = 1$ is a true constraint because it is equivalent to $\mathrm{var}(\mathbf{bX}) = 1$, which involves only the output $Y = \mathbf{bX}$ of the separation system (in the deflation approach).

To illustrate the difference between minimizing $H_r(Y)$, where $Y = \sum_{i=1}^{K} S_i$, under the constraint $\sum_{i=1}^{K} w_i^2 = 1$ and under the constraint $\max_{i=1}^{K} |w_i| = 1$, note that these minimizations are equivalent to minimizing without the constraint

$$H_r^*(Y) \doteq H_r\left[Y / \left(\sum_{i=1}^{K} w_i^2\right)^{1/2}\right] = H_r(Y) - \frac{1}{2}\log\left(\sum_{i=1}^{K} w_i^2\right)$$

in the first case and

$$H^\dagger(Y) = H_r(Y / \max_{i=1,\ldots,K} |w_i|) = H_r(Y) - \log \max_{i=1,\ldots,K} |w_i|$$

in the second case. One can see that $H^\dagger(Y)$ equals $H^*(Y)$ plus the term $(1/2)\log\left(\sum_{i=1}^{K} w_i^2 / \max_{j=1,\ldots,K} w_j^2\right)$, which reaches its (global) minimum (equal to zero) at and only at any nonmixing point (that is, a point for which all its coordinates except one are zero).[6] Therefore, even if $H^*(Y)$ does not admit a minimum at some nonmixing point, the addition of this term,

[6]To show that $\sum_{i=1}^{K} w_i^2 / \max_{j=1,\ldots,K} w_j^2$ reaches its minimum at and only at a nonmixing point, observe that w_i^2 equals $\max_{j=1,\ldots,K} w_j^2$ for at least an i, hence $\sum_{i=1}^{K} w_i^2 \geq \max_{j=1,\ldots,K} w_j^2$ with equality if and only if $w_i = 0$ for all except one index i

which does admit, could make this point a global minimum point of the sum (this is actually what happens). To illustrate further, consider the two-source case. Minimizing $H_r(Y)$ under the constraint $w_1^2 + w_2^2 = 1$ is the same as minimizing $H_r(Y_\theta)$, where $Y_\theta = (\sin\theta)S_1 + (\cos\theta)S_2$, with respect to θ. Minimizing $H_r(Y)$ under the constraint $\max(|w_1|, |w_2|) = 1$ is the same as minimizing $H_r[Y_\theta / \max(|\sin\theta|, |\cos\theta|)]$ with respect to θ. Because $H_r[Y_\theta / \max(|\sin\theta|, |\cos\theta|)] = H_r(Y_\theta) - \log\max(|\sin\theta|, |\cos\theta|)$, it is seen that changing the constraint $w_1^2 + w_2^2 = 1$ to $\max(|w_1|, |w_2|) = 1$ is equivalent to adding a term $-\log\max(|\sin\theta|, |\cos\theta|)$, which is minimized at and only at $\theta = k\pi/2$, which explains how a "noncontrast" function can be transformed into a "contrast" function by changing the constraint. Fig. 2 illustrates this phenomenon by plotting both $H_r(Y_\theta)$ (solid) and $H_r(Y_\theta) - \log\max(|\sin\theta|, |\cos\theta|)$ (dashed). The difference between the two curves represents the extra term, which was used to transform a "noncontrast" function to a "contrast" function. However, this term is *artificial* in the same way as the constraint $\max(|w_1|, |w_2|) = 1$ is, because both depend on the unknown mixing matrix.

Note that in the deconvolution case, Bercher and Vignat [2] circumvented the problem that $\max_{i=1,\ldots,K} |w_i| = 1$ is not a true constraint by fixing instead the value of the first coefficient of the deconvolution filter. This is equivalent to fixing the value of the first coefficient (instead of the *maximum absolute values of all coefficients*) of the global filter. This approach is, however, very specific to the deconvolution problem and not generalizable to the BSS case. Further, it suffers from the "nonrobustness" problem if the product of these first coefficients is small [2].

VI. Conclusion

Shannon's entropy has been proved to be a suitable functional to build contrast functions for BSS when at most one of the independent sources is Gaussian, and Hartley's one when the sources have finite support. Considering then the extended forms of Shannon's and Hartley's entropies, namely, Rényi's entropies, is thus appealing: for example, choosing $r = 2$ simplifies the computation of the entropy when the densities are estimated via Parzen window with Gaussian kernels [10]. Therefore, some authors have proposed to use the generalized Rényi's information measure, in particular, the quadratic entropy, to achieve BSS.

In this paper, a Taylor expansion of the Rényi's entropy has been performed and a sufficient condition that Rényi's entropy (with exponent not in $\{0, 1\}$) is not a contrast function has been given. Our results show that whatever the Rényi exponent $r \notin \{0, 1\}$ is, there always exists a non-Gaussian density such that the r-Rényi-entropy-based criterion is not a contrast function if the sources follow this density. Note that at points corresponding to satisfactory solution, it is shown that the criterion reaches a stationary point, but one does not know if they are maxima, minima, or even saddle points. This is a much harder situation than the one occurring with the kurtosis in the deflation approach, where the point corresponding to satisfactory solution is either a (global) minimum point or a maximum point, depending on the sub/super-Gaussianity of the sources, that is, on the sign of the kurtosis. In this case, the kurtosis criterion can be replaced by the absolute kurtosis (or any increasing function of it), which is then a contrast function. This is not applicable to Rényi-entropy-based criterion because one cannot infer the nature of the stationary point (local maxima, minima of saddle points) from the value of the criterion at this point. One may try to estimate the rth Fisher information and check the condition of Theorem 2, but this condition only implies that the above stationary point is not a local maxima point: it can still be a saddle or local (but not global) minima point so that changing the sign of the criterion does not really help. Another possibility is to estimate the "good choice" of r from the data, but this implies that there exists a simple rule relating such "good choice" to the nature of the sources (sur/sub Gaussian, for example), which, to our best knowledge, does not seem to exist. Another problem with the above approaches is that the source characteristics have to be estimated from the extracted sources, but if the criterion is not a contrast function, these extracted sources may be still a mixture, hence the estimates would be wrong.

This means that using the general form of the Rényi's entropy may be risky. The Rényi's exponent $r = 1$ (Shannon's entropy) and $r = 0$ (support measure) cases seem thus to have a very specific behavior in the context of BSS. As mentioned in the Introduction, only Shannon's and Hartley's entropies possess the superadditivity property (because this property would imply the contrast function property of the corresponding BSS criterion). Interestingly, other authors, in somewhat different contexts, have proved that these two entropies indeed possess specific properties, not shared by the other entropies belonging to the Rényi's family [1], [4], [27].

Appendix

A. First-Order Taylor Expansion of Rényi's Entropy

We provide here the first-order Taylor expansion of the Rényi entropy, a random Y contaminated by a small random variable of the form $\mathbf{hZ}$ where $\mathbf{h}$ is a small row vector and $\mathbf{Z}$ is a column random vector. It is proved in [22, Lemmas 1 and 3] that under appropriate conditions, the density $p_{Y+\mathbf{hZ}}$ of $Y+\mathbf{hZ}$ admits the continuous partial derivatives with respect to the components of $\mathbf{h}$, with the vector of the derivatives $\partial p_{Y+\mathbf{hZ}}/\partial\mathbf{h}$ satisfying

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\partial p_{Y+\mathbf{hZ}}(y)}{\partial\mathbf{h}} = -\frac{d[E(\mathbf{Z}|Y=y)p_Y(y)]}{dy} \qquad \forall\, y \quad (25)$$

where $E(\mathbf{Z}|Y=y)$ is the conditional expectation of $\mathbf{Z}$ given $Y=y$ and p_Y is the density of Y. On the other hand, assuming that one can interchange the order of differentiation and integration

$$\frac{\partial}{\partial\mathbf{h}} \int p_{Y+\mathbf{hZ}}^r(y)dy = r\int \frac{\partial p_{Y+\mathbf{hZ}}(y)}{\partial\mathbf{h}} p_{Y+\mathbf{hZ}}^{r-1}(y)dy. \quad (26)$$

Such interchange can be justified (via the Fubini theorem [28]) if $\int \|\partial p_{Y+\mathbf{hZ}}(y)/\partial\mathbf{h}\| p_{Y+\mathbf{hZ}}^{r-1}(y)dy$ exists and is bounded for all $\mathbf{h}$ small enough.[7] If, moreover, the function under this integral sign can be bounded for all $\mathbf{h}$ small enough by an inte-

[7]Such condition can be fulfilled if the joint density of Y and $\mathbf{Z}$ and its partial derivative with respect to the first argument go to 0 sufficiently fast at infinity.

grable function, then by the Lebesgue dominated convergence theorem, (25) and (26) yield

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\partial}{\partial \mathbf{h}} \int p^r_{Y+\mathbf{hZ}}(y)dy = -r \int \frac{d[E(\mathbf{Z}|Y=y)p_Y(y)]}{dy} p^{r-1}_Y(y)dy$$

which equals $r(r-1)\int E(\mathbf{Z}|Y=y)p^{r-1}_Y(y)p'_Y(y)dy$ by integration by parts, $'$ denoting the derivative. It then follows from the definition of Rényi's entropy given in (4) that $H_r(Y+\mathbf{hZ})$, as a function of $\mathbf{h}$, is continuously differentiable with the vector of the derivative satisfying

$$\begin{aligned}\lim_{\mathbf{h}\to\mathbf{0}} &\frac{\partial}{\partial \mathbf{h}} H_r(Y+\mathbf{hZ}) \\ &= -r\frac{\int E(\mathbf{Z}|Y=y)p^{r-1}_Y(y)p'_Y(y)(y)dy}{\int p^r_Y(y)dy} \\ &= -E\left[\frac{rp^{r-2}_Y(Y)p'_Y(Y)}{\int p^r_Y(y)dy}\mathbf{Z}\right].\end{aligned}$$

This yields the expansion given in (5).

B. Second-Order Taylor Expansion of Rényi's Entropy

We compute here the second-order expansion of $H_r(Y+\mathbf{hZ})$ *in the case where* $\mathbf{Z}$ *is independent of* Y. For convenient, we will assume that $\mathbf{Z}$ has zero mean. This does not affect the generality because $H_r(Y+\mathbf{hZ})$ is unchanged when one subtracts its $\mathbf{Z}$ mean. Then, $\lim_{\mathbf{h}\to\mathbf{0}} \partial p_{Y+\mathbf{hZ}}/\partial \mathbf{h} = \mathbf{0}$ by (25) and the fact that $E(\mathbf{Z}|Y=y) = E(\mathbf{Z}) = 0$. Hence

$$\frac{\partial p_{Y+\mathbf{hZ}}(y)}{\partial \mathbf{h}} p^{r-1}_{Y+\mathbf{hZ}}(y) \quad \text{and} \quad \frac{\partial H(Y+\mathbf{hZ})}{\partial \mathbf{h}}$$

both tend to $\mathbf{0}$ as $\mathbf{h}\to\mathbf{0}$.[8]

From the results of [22], $p_{Y+\mathbf{hZ}}$ admits (under appropriate assumptions) the second partial derivatives with respect to the components of $\mathbf{h}$, with the matrix of the derivatives $\partial^2 p_{Y+\mathbf{hZ}}/\partial \mathbf{h}^2$ satisfying

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\partial^2 p_{Y+\mathbf{hZ}}(y)}{\partial \mathbf{h}^2} = E(\mathbf{ZZ}^T)p''_Y(y) \qquad \forall\, y \qquad (27)$$

where $''$ denotes the second derivative. On the other hand, taking the partial derivatives of both sides of (26) and assuming again that one can interchange the order of differentiation and integration, one gets

$$\begin{aligned}\frac{\partial^2}{\partial \mathbf{h}^2} &\int p^r_{Y+\mathbf{hZ}}(y)dy \\ &= r\int \left[\frac{\partial^2 p_{Y+\mathbf{hZ}}(y)}{\partial \mathbf{h}^2} p^{r-1}_{Y+\mathbf{hZ}}(y) + (r-1)\right. \\ &\qquad \left.\times \frac{\partial p_{Y+\mathbf{hZ}}(y)}{\partial \mathbf{h}} \frac{\partial p_{Y+\mathbf{hZ}}(y)}{\partial \mathbf{h}}^T p^{r-2}_{Y+\mathbf{hZ}}(y)\right] dy.\end{aligned}$$

Again, such interchange can be justified if the integrals $\int \|\partial^2 p_{Y+\mathbf{hZ}}(y)/\partial \mathbf{h}^2\| p^{r-1}_{Y+\mathbf{hZ}}(y)dy$ and $\int \|\partial p_{Y+\mathbf{hZ}}(y)/\partial \mathbf{h}\|^2 p^{r-2}_{Y+\mathbf{hZ}}(y)dy$ exist and are bounded for all $\mathbf{h}$ small enough. If, moreover, the functions under these integral signs can be bounded for all $\mathbf{h}$ small enough by some integrable functions, then again by the Lebesgue dominated convergence theorem, one gets

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\partial^2}{\partial \mathbf{h}^2} \int p^r_{Y+\mathbf{hZ}}(y)dy = rE(\mathbf{ZZ}^T) \int p''_Y(y)p^{r-1}_Y(y)dy$$

because $\lim_{\mathbf{h}\to\mathbf{0}} \partial p_{Y+\mathbf{hZ}}/\partial \mathbf{h} = \mathbf{0}$.

Finally

$$\begin{aligned}&\frac{\partial^2}{\partial \mathbf{h}^2} \log \int p^r_{Y+\mathbf{hZ}}(y)dy \\ &\quad = \frac{1}{\int p^r_{Y+\mathbf{hZ}}(y)dy} \frac{\partial^2 \int p^r_{Y+\mathbf{hZ}}(y)dy}{\partial \mathbf{h}^2} \\ &\qquad - \frac{\partial}{\partial \mathbf{h}} \log \int p^r_{Y+\mathbf{hZ}}(y)dy \left[\frac{\partial}{\partial \mathbf{h}} \log \int p^r_{Y+\mathbf{hZ}}(y)dy\right]^T\end{aligned}$$

and because $\partial H(Y+\mathbf{hZ})/\partial \mathbf{h}$ tends to $\mathbf{0}$ as $\mathbf{h}\to\mathbf{0}$, by (4), the same holds for the last term in the above right-hand side and

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\partial^2}{\partial \mathbf{h}^2} H(Y+\mathbf{hZ}) = \frac{r}{1-r} \frac{\int p''_Y(y)p^{r-1}_Y(y)dy}{\int p^r_Y(y)dy} E(\mathbf{ZZ}^T).$$

Noting that $E(\mathbf{ZZ}^T)$ is also the covariance matrix of $\mathbf{Z}$, one gets the second-order Taylor expansion given in (12). As both sides of (12) remain unchanged when one adds a constant vector to $\mathbf{Z}$, this formula is still valid for the noncentered random variables $\mathbf{Z}$ (but independent of Y).

[8]Actually, the convergence to $\mathbf{0}$ of $\partial H(Y+\mathbf{hZ})/\partial \mathbf{h}$ does not require that Z has zero mean because the Rényi's entropy is translation invariant.

REFERENCES

[1] J. Aczel, B. Forte, and C. T. Ng, "Why the Shannon and Hartley entropies are natural ?," *Adv. Appl. Probab.*, vol. 6, pp. 131–146, 1974.

[2] J.-F. Bercher and C. Vignat, "A Rényi entropy convolution inequality with application," in *Proc. Eur. Signal Process. Conf.*, Toulouse, France, 2002, vol. 2, pp. 111–114.

[3] P. Comon, "Independent component analysis, a new concept?," *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.

[4] M. Costa and T. Cover, "On the similarity of the entropy power inequality and the Brunn-Minkowski inequality," *IEEE Trans. Inf. Theory*, vol. IT-30, no. 6, pp. 837–839, Nov. 1984.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, ser. Telecommunications. New York: Wiley, 1991.

[6] S. Cruces, A. Cichocki, and S. Amari, "From blind signal extraction to blind instantaneous signal separation: Criteria, algorithms and stability," *IEEE Trans. Neural Netw.*, vol. 15, no. 4, pp. 859–873, Jul. 2004.

[7] S. Cruces and I. Duran, "The minimum support criterion for blind source extraction: A limiting case of the strengthened young's inequality," in *Lecture Notes in Computer Science*, ser. 3195, C. G. Puntonet and A. Prieto, Eds. Berlin, Germany: Springer-Verlag, Sep. 2004, pp. 57–64.

[8] D. Erdogmus, K. E. Hild, and J. Principe, "Blind source separation using Rényi's α -marginal entropies," *Neurocomputing*, vol. 49, no. 49, pp. 25–38, 2002.

[9] D. Erdogmus, K. E. Hild, and J. Principe, "Correction to blind source separation using Rényi's mutual information," *IEEE Signal Process. Lett.*, vol. 10, no. 8, p. 250, Aug. 2003.

[10] D. Erdogmus, K. E. Hild, II, M. Lazaro, I. Santamaria, and J. C. Principe, "Adaptive blind deconvolution of linear channels using Rényi's entropy with Parzen estimation," *IEEE Trans. Signal Process.*, vol. 52, no. 6, pp. 1489–1498, Jun. 2004.

[11] D. Erdogmus, J. C. Principe, and L. Vielva, "Blind deconvolution with Rényi's minimum entropy," in *Proc. Eur. Signal Process. Conf.*, Toulouse, France, 2002, vol. 1, pp. 557–560.

[12] R. J. Gardner, "The Brunn-Minkowski inequality," *Bull. Amer. Math. Soc.*, vol. 3, no. 39, pp. 355–405, 2002.

[13] K. E. Hild, D. Erdogmus, and J. Principe, "Blind source separation using Rényi's mutual information," *IEEE Signal Process. Lett.*, vol. 8, no. 6, pp. 174–176, Jun. 2001.

[14] K. E. Hild, D. Erdogmus, and J. Principe, "An analysis of entropy estimators for blind source separation," *Signal Process.*, vol. 86, pp. 174-176182–194, 2006.
[15] P. J. Huber, "Projection pursuit," *Ann. Statist.*, vol. 13, no. 2, pp. 435–475, 1985.
[16] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.
[17] K. H. Knuth, "Lattice duality: The origin of probability and entropy," *Neurocomputing*, vol. 67, pp. 245–274, 2005.
[18] E. Lutwak, D. Yang, and G. Zhang, "Cramér-Rao and moment-entropy inequalities for Rényi entropy and generalized Fisher information," *IEEE Trans. Inf. Theory*, vol. 51, no. 2, pp. 473–478, Feb. 2005.
[19] D.-T. Pham, "Blind separation of instantaneous mixtures of sources based on order statistics," *IEEE Trans. Signal Process.*, vol. 48, no. 2, pp. 363–375, Feb. 2000.
[20] D.-T. Pham, "Contrast functions for blind separation and deconvolution of sources," in *Proc. Int. Conf. Independ. Component Anal. Blind Signal Separat.*, T.W. Lee, T. P. Jung, S. Makeig, and T. J. Sejnowski, Eds., San Diego, CA, 2001, pp. 37–42.
[21] D.-T. Pham, "Mutual information approach to blind separation of stationary sources," *IEEE Trans. Inf. Theory*, vol. 48, no. 7, pp. 1935–1946, Jul. 2002.
[22] D.-T. Pham, "Entropy of a variable slightly contaminated with another," *IEEE Signal Process. Lett.*, vol. 12, no. 7, pp. 536–539, Jul. 2005.
[23] J. C. Principe and D. Xu, "Information theoretic learning using Rényi quadratic entropy," in *Proc. Int. Conf. Independ. Component Anal. Blind Signal Separat.*, J. -C. Cardoso, C. Jutten, and P. Loubaton, Eds., Aussois, France, Jan. 1999, pp. 407–412.
[24] J. C. Principe, D. Xu, and J. W. Fisher, III, "Information-theoretic learning," in *Unsupervised Adaptive Filtering*, ser. Adaptive Learning Systems for Signal Processing, Communications, and Control, S. Haykin, Ed. New York: Wiley, 2000, vol. 1, pp. 265–319.
[25] A. Renyi, *Calcul des Probabilités*. Paris, France: Dunod, 1966.
[26] A. Renyi, "On measures of entropy and information," *Selected Papers of Alfred Renyi*, vol. 2, pp. 565–580, 1976.
[27] A. Renyi, "Some fundamentals questions of information theory," *Selected Papers of Alfred Renyi*, vol. 2, pp. 526–552, 1976.
[28] W. Rudin, *Real and Complex Analysis*. New York: McGraw-Hill, 1987.
[29] F. Vrins, J. A. Lee, and M. Verleysen, "A minimum-range approach to blind extraction of bounded sources," *IEEE Trans. Neural Netw.*, vol. 18, no. 3, pp. 809–822, May 2007.
[30] F. Vrins and D.-T. Pham, "Minimum range approach to blind partial simultaneous separation of bounded sources: Contrast and discriminacy properties," *Neurocomputing*, vol. 70, no. 7–9, pp. 1207–1214, 2007.

Dinh-Tuan Pham (M'88) was born in Hanoï, Vietnam, on February 10, 1945. He graduated from the Engineering School of Applied Mathematics and Computer Science (ENSIMAG), Polytechnic Institute of Grenoble, Grenoble, France, in 1968. He received the Ph.D. degree in statistics from the University of Grenoble, Grenoble, France, in 1975.

He was a Postdoctoral Fellow at the Department of Statistics, University of California, Berkeley, in 1977–1978 and a Visiting Professor at the Department of Mathematics, Indiana University, Bloomington, in 1979–1980. Currently, he is Director of Research at the French Centre National de Recherche Scientifique (CBRS), Laboratoire Jean Kuntzmann. His research interests include time-series analysis, signal modeling, blind source separation, nonlinear (particle) filtering, and biomedical signal processing.

Frédéric Vrins (S'06) was born in Uccle, Belgium, in 1979. He received the M.S. degree in mechatronics engineering and the DEA and Ph.D. degreed in applied sciences from the Université catholique de Louvain, Louvain-la-Neuve, Belgium, in 2002, 2004, and 2007, respectively.

His research interests are blind source separation, independent component analysis, Shannon and Rényi entropies, mutual information, and information theory in adaptive signal processing. Currently, he is a Quantitative Analyst (Modeler) at the Financial Markets Department, ING Wholesale Banking, Brussels, Belgium.

Michel Verleysen (S'87–M'92–SM'04) was born in Belgium in 1965. He received the M.S. and Ph.D. degrees in electrical engineering from the Université catholique de Louvain, Louvain-la-Neuve, Belgium, in 1987 and 1992, respectively.

He was an Invited Professor at the Swiss Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, in 1992, at the Université d'Evry Val d'Essonne, France, in 2001, and at the Université Paris I-Panthéon-Sorbonne, Paris, France, in 2002, 2003, and 2004. Currently, he is Research Director of the Belgian Fonds National de la Recherche Scientique (FNRS) and Lecturer at the Université catholique de Louvain. He is author or coauthor of about 200 scientific papers in international journals and books or communications to conferences with reviewing committee. He is the coauthor of the scientific popularization book on artificial neural networks in the series *Que Sais-Je?* (France: Presses Universitaires de France, 1996, in French). His research interests are in artificial neural networks, self-organization, time-series forecasting, nonlinear statistics, adaptive signal processing, information-theoretic learning and biomedical data, and signal analysis.

He is Editor-in-Chief of the *Neural Processing Letters* and Chairman of the annual European Symposium on Artificial Neural Networks (ESANN) conference. He is Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS, and member of the editorial board and program committee of several journals and conferences on neural networks and learning.