Information Theoretic versus Cumulant-Based Contrasts for Multimodal Source Separation

Frédéric Vrins and Michel Verleysen, *Senior Member, IEEE*

***Abstract*—Recently, several authors have emphasized the existence of spurious maxima in usual contrast functions for source separation (e.g., the likelihood and the mutual information) when several sources have multimodal distributions. The aim of this letter is to compare the information theoretic contrasts to cumulant-based ones from the robustness to spurious maxima point of view. Even if all of them tend to measure, in some way, the same quantity, which is the output independence (or equivalently, the output non-Gaussianity), it is shown that in the case of a mixture involving two sources, the kurtosis-based contrast functions are more robust than the information theoretic ones when the source distributions are multimodal.**

***Index Terms*—Blind source separation, contrast function, entropy, independent component analysis, kurtosis, multimodal sources.**

I. Introduction

Blind source separation (BSS) consists in recovering independent source signals $\mathbf{s}(t) = [s_1(t), \ldots, s_m(t)]$ from n mixtures of them $x_i(t) = \mathcal{A}_i(\mathbf{s}(t))$ $(1 \leq i \leq n)$. In this letter, we focus on the linear instantaneous mixture of real sources $\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$, where $\mathbf{x} = [x_1, \ldots, x_n]$, and $\mathbf{A}$ denotes the mixing matrix (with a slight abuse of notation, we will omit the temporal variable t in the following). At most, one source may have a normal distribution. The mixing system is supposed to be square $(m = n)$. Without loss of generality (provided that the sources are stationary and ergodic), it is commonly assumed that the sources are zero-mean and have an identity covariance matrix (i.e., they are *sphered*).

Most of time, the data $\mathbf{x}$ are sphered using a *prewhitening* step: $\mathbf{x}' \leftarrow \mathbf{W}\mathbf{x}$, such that $E\{\mathbf{x}'\} = \mathbf{0}$ and $E\{\mathbf{x}'\mathbf{x}'^T\} = \mathbf{I}_n$. If we furthermore constrain the estimated sources $\mathbf{y}$ (also called "output signals") to be sphered, they become a rotation transform of $\mathbf{x}'$. If $\mathbf{U}$ symbolizes the rotation matrix, the mixture scheme can be rewritten as

$$\mathbf{y} = \mathbf{U}\underbrace{\mathbf{W}\mathbf{x}}_{\mathbf{x}'} = \underbrace{\mathbf{U}\mathbf{W}\mathbf{A}}_{\mathbf{C}}\mathbf{s} \tag{1}$$

where $\mathbf{C}$ denotes the *transfer matrix* between the outputs and the source signals. The aim of BSS is to obtain output signals $\mathbf{y}$ that correspond to the original sources. In this case, the square matrix solution $\mathbf{C}^\star \doteq \mathbf{U}^\star\mathbf{W}\mathbf{A}$ is nonmixing (at most, one nonzero element per row and full rank) [1]; matrix $\mathbf{U}^\star$ is the rotation matrix, maximizing a so-called contrast function Φ, i.e., $\mathbf{U}^\star = \arg\max_{\mathbf{U}} \Phi$. When independent component analysis (ICA) is used to solve the BSS problem, Φ is a function that measures the independence level between the elements of $\mathbf{y}$ [1]. In order to avoid an exhaustive search in the whole space of orthogonal matrices, a gradient ascent on Φ is used most of the time, leading to an update rule for $\mathbf{U}$ that looks like

$$\mathbf{U}(k+1) \leftarrow \mathbf{U}(k) + \mu(k)\nabla\Phi|_{\mathbf{U}(k)} \tag{2}$$

In (2), $\nabla\Phi|_{\mathbf{U}(k)}$ may denote either the Euclidean, natural, or relative gradient of Φ with respect to $\mathbf{U}$, evaluated at $\mathbf{U} = \mathbf{U}(k)$. Note that algebraic methods also exist for specific contrast functions Φ.

Using a gradient-based maximization supposes that the algorithm will not be trapped in a spurious maximum, leading to $\widetilde{\mathbf{U}}$, that does not correspond to a satisfactory solution for the BSS problem ($\widetilde{\mathbf{C}} = \widetilde{\mathbf{U}}\mathbf{W}\mathbf{A}$ still mixing). In [2]–[6], various authors have noted that the usual ICA contrast functions may have such spurious maxima if several source distributions are multimodal. For instance, Cardoso shows this phenomenon in [6] for the likelihood-based contrast function Φ_{ML} and explains it as a local matching between the distribution $p_{\mathbf{y}}(\mathbf{y})$ of $\mathbf{y}$ and the supposed distribution $\widehat{p}(\mathbf{s})$ used in Φ_{ML}, even if a correct model has been assumed for the source distributions, i.e., even if $\widehat{p}(\mathbf{s}) = p_{\mathbf{s}}(\mathbf{s})$. More recently, Vrins *et al.* [3], [4] have given an intuitive justification regarding the existence of spurious maxima when the opposite of the output marginal entropies are used for the contrast function. This can be understood looking at the structure of the $p_{y_i}(y_i)$ and, more precisely, their number of modes $N(y_i)$ (see Section II for a summary of these results).

This paper aims to show that cumulant-based contrast functions do not suffer from this drawback, at least if $n = 2$. After analyzing the robustness of the entropic contrasts to the existence of spurious maxima, we justify the use of the kurtosis as the contrast function to separate two multimodal sources.

II. Entropy Spurious Minima

Consider a two inputs and two outputs (TITO) system $(n = m = 2)$. The transfer matrix $\mathbf{C}$ can be modeled as in

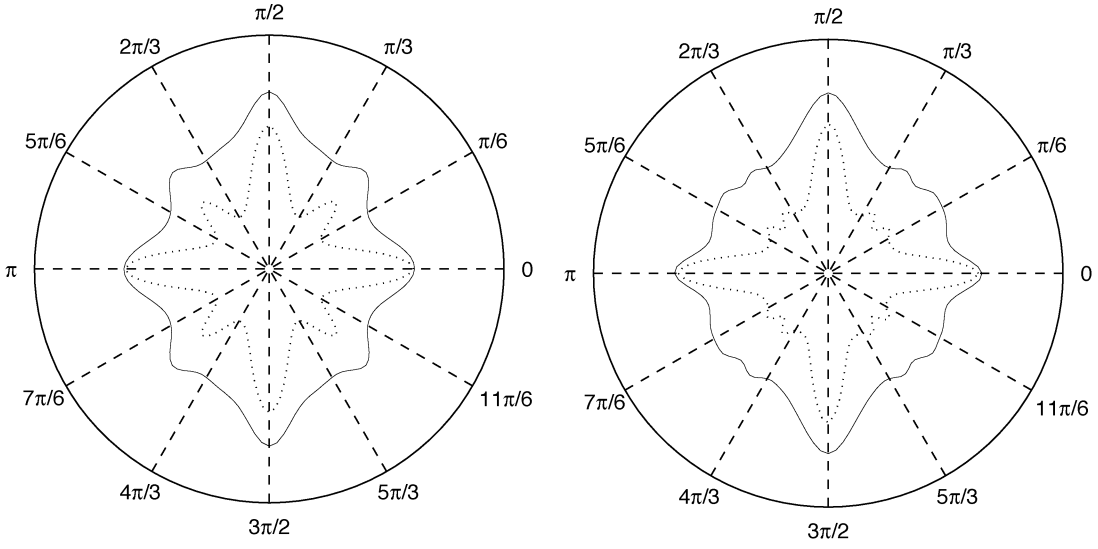
$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \sin\theta & \cos\theta \\ -\cos\theta & \sin\theta \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix}. \tag{3}$$

This particular form of $\mathbf{C}$ is due to the fact that i) both matrices $\mathbf{U}$ and $\mathbf{W}\mathbf{A}$ are orthogonal (so that $\mathbf{C}$ is also orthogonal)

Manuscript received July 5, 2004; revised September 7, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yiteng (Arden) Huang.

The authors are with the UCL Machine Learning Group, Université Catholique de Louvain, Louvain-la-Neuve, Belgium (e-mail: vrins@dice.ucl.ac.be).

Digital Object Identifier 10.1109/LSP.2004.840863



(a) $\widehat{\Phi}_{H1}$ (solid) and $\widehat{\Phi}_H$ (dotted) vs θ ($N(s_1) = N(s_2) = 2$).

(b) $\widehat{\Phi}_{H1}$ (solid) and $\widehat{\Phi}_H$ (dotted) vs θ ($N(s_1) = 2$, $N(s_2) = 3$).

Fig. 1. Spurious maxima of $\widehat{\Phi}_{H1}$ and $\widehat{\Phi}_H$ (corresponding to $\theta \in]k\pi/2, (k+1)\pi/2[$) for (a) a pair of bimodal sources and for (b) a mixture of a bimodal and a trimodal sources.

and ii) in dimension two, an orthogonal matrix is fully determined by a single angle. Since, in practice, $\mathbf{A}$ is unknown, the angle θ (which is a function of the elements of $\mathbf{A}$, $\mathbf{W}$, and $\mathbf{U}$) is unknown, too. However, the angle θ may be *blindly* modified through the elements of $\mathbf{U}$.

Obviously, all unmixing matrices $\mathbf{U}^\star$ corresponding to $\theta = k\pi/2$ ($k \in \mathbb{Z}$) are acceptable solutions for the BSS problem, since they are associated to nonmixing matrices $\mathbf{C}^\star$. We will focus on $\Phi_{H1} \doteq -H(y_1)$ and $\Phi_H \doteq -\sum_{i=1}^n H(y_i)$, where $H(y) = -\int p_y(\xi) \log p_y(\xi) d\xi$ denotes the entropy of y [8]. These latter criteria can be used as a contrast function for ICA [7] (of course, they do not involve the unknown part of the mixing model: neither the elements of $\mathbf{A}$ nor the unknown sources s_i). The choice between these two criteria depends on if a deflationist approach (the sources are estimated one by one) or a symmetric one (both sources are extracted simultaneously) is preferred. Note that the exact computation of the entropy $H(y_1)$ requires that you know the distribution $p_{y_1}(y_1)$ of the variable y_1. Since, in practice, the latter is unknown, the distributions will be estimated, for example, using the Parzen estimator [9] with Gaussian kernels (see [4] for more details about the choice of the kernel variance in this application). Fig. 1 shows the evolution of $\widehat{\Phi}_{H1} \doteq \zeta + \Phi_{H1}$ and $\widehat{\Phi}_H \doteq \zeta + \Phi_H$ versus θ for two examples (ζ is a well-chosen scalar, ensuring that $\widehat{\Phi}_H$ and $\widehat{\Phi}_{H1}$ are positive, for illustration purposes). In Fig. 1(a), both source distributions are bimodal: $N(s_1) = N(s_2) = 2$, while in Fig. 1(b), $N(s_1) = 2$ and $N(s_2) = 3$ (since the scale of the axes does not matter, it has been omitted). The distributions $p_{s_i}(s_i)$ are built by adding $N(s_i)$ Gaussian kernels of different means (with negligible overlap). Keeping in mind that the only maxima that are relevant from the BSS point of view are the ones occurring at $\theta = \{\pi/2 | k \in Z\}$, it is obvious that both these contrast functions have spurious maxima. A BSS algorithm using (2) may fail in such cases.

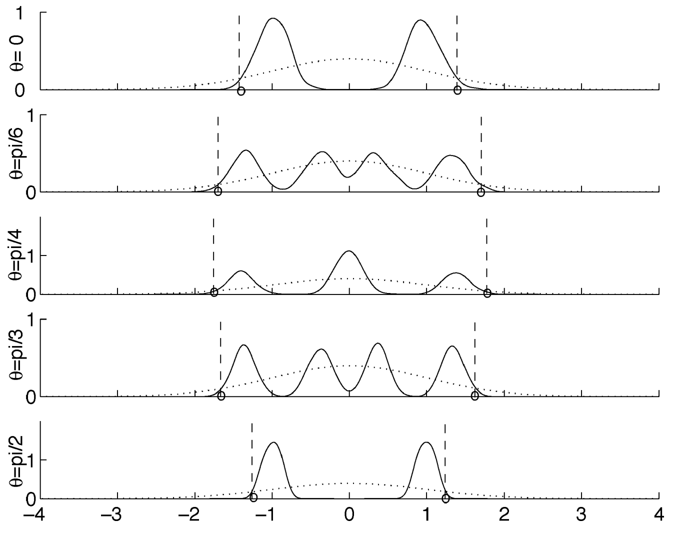
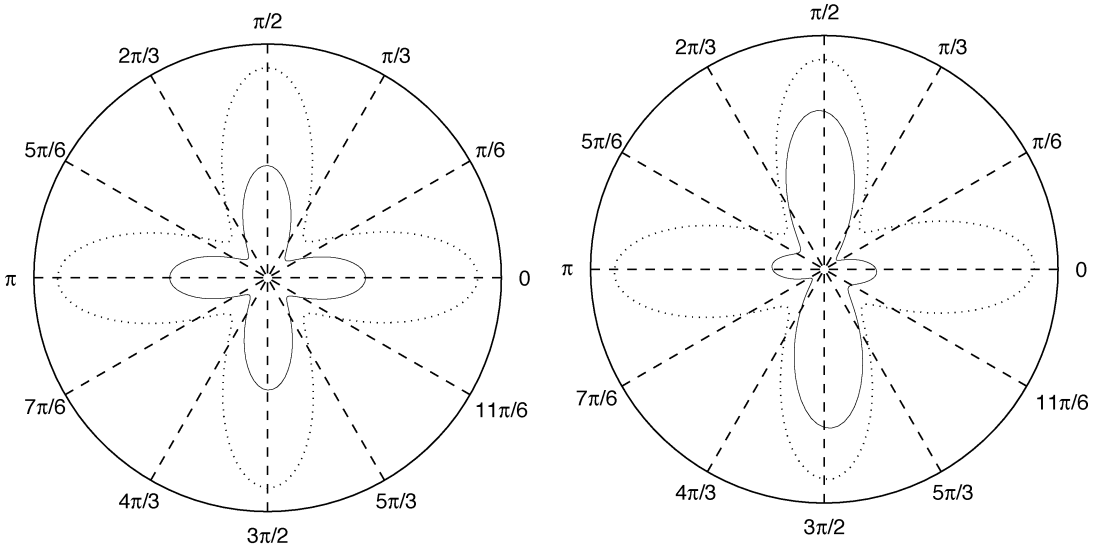


Fig. 2. Evolution of $p_{y_1}(y_1)$ for several values of θ (solid) associated with Fig. 1(a) and the sphered Normal distribution (dotted).

In [4], it is explained that $N(y_1)$ may vary between $\min(N(s_1), N(s_2))$ and $N(s_1).N(s_2)$ for θ varying between $[k\pi/2, (k+1)\pi/2]$. It is emphasized that $N(y_1)$, when expressed as a function of θ, may have local minima in $]k\pi/2, (k+1)\pi/2[$; these minima coincide with the (spurious) local minima of $H(y_1)$, i.e., the spurious local maxima of Φ_{H1}. This can be observed by comparing Figs. 1(a) and 2. This phenomenon is due to the fact that the distribution of a sum of independent random variables is the convolution of the variable distributions (see Section IV).

The analysis in this section has been extended to nearest-neighbor approximators of entropy (spacing estimates of entropy [5]) and to other definitions of entropy (like Renyi's entropy [10]): The conclusion is identical.



(a) $\Phi_{\kappa 1}$ (solid) and Φ_{κ} (dotted) vs θ ($N(s_1) = N(s_2) = 2$).

(b) $\Phi_{\kappa 1}$ (solid) and Φ_{κ} (dotted) vs θ ($N(s_1) = 2$, $N(s_2) = 3$).

Fig. 3. Evolution of $\Phi_{\kappa 1}$ and Φ_{κ} for (a) a pair of bimodal sources and for (b) a mixture of a bimodal and a trimodal sources.

Since, in this mixture, scheme Φ_H is equivalent to the opposite of the mutual information and to the negentropy criteria [11], the latter obviously suffers from the same drawback.

III. Kurtosis-Based Contrast Function

As in the previous section, similar simulations were performed, now using the absolute value of the kurtosis as the contrast function instead of Shannon's entropy. This fourth-order cumulant can be used in BSS applications (see [12] and inner references). Hence, the aim of ICA is to find the $\mathbf{U}^\star$ that maximizes $\Phi_{\kappa 1} \doteq |\kappa(y_1)|$ or $\Phi_\kappa \doteq |\kappa(y_1)| + |\kappa(y_2)|$ (both can be used, depending on if a deflationist or a symmetric approach is used for the separation). The results are plotted in Fig. 3. As in the previous section, the values of θ corresponding to nonmixing matrices $\mathbf{C}^\star$ are $\{k\pi/2\}$. We can observe that all the local maxima of $\Phi_{\kappa 1}$ or Φ_κ correspond to a nonmixing matrix $\mathbf{C}^\star$.

IV. Discussion

In this section, we compare the entropy- and kurtosis-based contrast functions from the viewpoint of spurious maxima. information theoretic criteria, as well as cumulant-based ones, map the structure of a distribution to a real number. Both these criteria measure statistical quantities of distributions.

The distribution of y_1 is directly related to $p_{s_1}(s_1)$, $p_{s_2}(s_2)$, and θ. Indeed, multiplying a variable u is equivalent to scaling $p_u(u)$:

$$p_{\alpha u}(\alpha u) = \frac{1}{|\alpha|} p_u(u) \ (\alpha \in \mathbb{R}) \tag{4}$$

and the distribution that results from the sum of independent random variables is the convolution of the variable distributions:

$$p_{u+v}(\xi) = \int_{-\infty}^{+\infty} p_u(\tau) p_v(\xi - \tau) d\tau. \tag{5}$$

Comparing Figs. 1(a) and 2, it is clear that Φ_{H1} is a measure of the whole structure of $p_{y_1}(y_1)$ (and, thus, depends on the number of modes). By contrast, $\Phi_{\kappa 1}$ characterizes more specifically the tails of $p_{y_1}(y_1)$, discarding its internal structure (in the middle range of the support of y_1), as is visible by comparing Figs. 2 to 3(a).

This property of the kurtosis $\kappa(y_1)$, which can be used as a non-Gaussianity measure of $p_{y_1}(y_1)$ [11], has been emphasized by Friedman: $(\ldots)$ *projection indexes based on standardized cumulants heavily emphasize the departure from normality in the tails of distribution.* $(\ldots)$ *For example, a distribution with only slightly heavier than normal tails receives a much higher index value than a highly clustered projection* (i.e., distribution) [13]. This analysis (particularized to the kurtosis) has been translated for the ICA problem in [14].

The previous considerations are illustrated in Figs. 2, 3(a), and 4. In this experience, $\kappa(y_1)$ can be seen approximately as a measure of *where the tails of* $p_{y_1}(y_1)$ *cross the tails of the Gaussian distribution* G *of zero mean and unit variance*. In other words, if we suppose that $G(y_1) \geq p_{y_1}(y_1)$ for $y_1 \geq y_1^{\star r}$ and for $y_1 \leq y_1^{\star l}$ (with $y_1^{\star l} < 0 < y_1^{\star r}$), then the lower $y_1^{\star r}$ and $|y_1^{\star l}|$, the higher $|\kappa(y_1)|$ [the link between the kurtosis and $|y_1^\star|$ can be seen by comparing Figs. 3(a) and 4]. The $y_1^{\star r}$ and $|y_1^{\star l}|$ are indicated by circles in Fig. 2. Note that $y_1^{\star r}$ or $|y_1^{\star l}|$ have similar behavior versus θ. Moreover, as visible in Fig. 4, $|y_1^{\star l}(\theta)| = y_1^{\star r}(\theta + \pi)$ and $y_1^{\star r}(\theta) = |y_1^{\star l}(\theta + \pi)|$ [since, by (3), $y_i(\theta) = -y_i(\theta + \pi)$]. Hence, the lower the $y_1^\star \doteq (y_1^{\star r} + |y_1^{\star l}|)/2$, the higher $|\kappa(y_1)|$.

The key point here is to observe that the evolution of $H(y_i)$ is largely influenced by $N(s_1)$ and $N(s_2)$. On the contrary, the evolution of $y_1^\star$ versus θ (or, more precisely, the shape of this function) mainly depends on the transfer coefficients (i.e., of θ); the number $N(y_1)$ of modes has no influence on the number of extrema of the kurtosis. Even if the source distributions stretch or distort the shape of $|\kappa(y_1)|$ (expressed as a function of θ), this shape remains similar for both unimodal or multimodal source distributions.

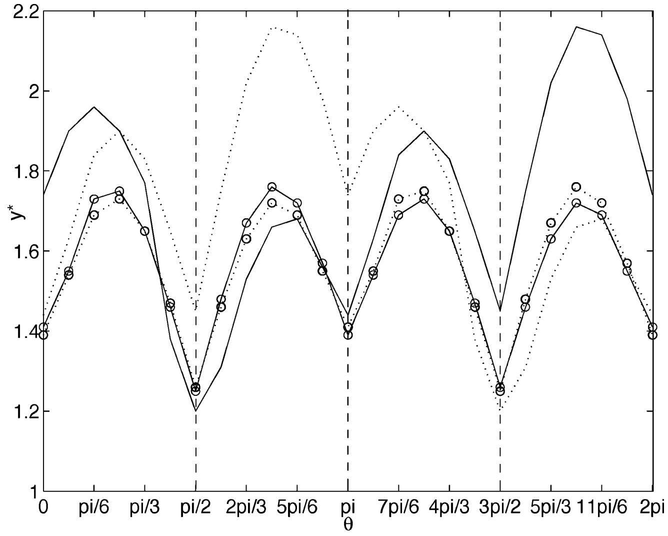


Fig. 4. Evolution of $y_1^{\star r}$ (solid) and $|y_1^{\star l}|$ (dotted) versus θ for the examples given in Fig. 3(a) (markers "o") and (b) (no marker).

Let us define $s_i^{\star r}$, $s_i^{\star l}$, and $s_i^\star$ similarly to $y_1^{\star r}$, $y_1^{\star l}$, and $y_1^\star$, respectively). Consequently, starting from $\theta = k\pi$ to $\theta = (2k+1)\pi/2$, $y_1^\star$ increases from $s_2^\star$, reaches a (possibly locally) maximum value, and decreases to $s_1^\star$. This is exactly the same scheme as for unimodal source separation, and it ensures that all locally maximum values of $|\kappa(y_1)|$ (i.e., the minimum values of $y_1^\star$), which can be detected blindly knowing only $\mathbf{U}$, are attained for $\theta = \{k\pi/2\}|(k \in \mathbb{Z})$, corresponding to a nonmixing transfer matrix $\mathbf{C}^\star$. As a consequence, using gradient-based maximization of Φ_{κ_1} or Φ_κ does not lead to spurious solutions. In addition, it is known that algebraic methods can also be used to maximize the last contrast function [15], avoiding spurious solutions, too. On the contrary, entropy-based contrast functions are maximized by gradient-based methods; it is shown in this paper that spurious maxima may appear in this case.

V. Conclusion

In the ICA community, despite its extreme simplicity, the kurtosis-based contrast functions are criticized for their low robustness to outliers. However, this behavior may constitute an advantage in some situations, as in the problem exposed here: It allows the characterization of the tails of a distribution, discarding the internal structure (in the middle range of the support of the variable). This is exactly what is desired when the goal is to separate multimodal sources, since the tails of $p_{y_1}(y_1)$ do not depend on $N(y_1)$, but rather only on the source distribution *tails* and the transfer coefficients (elements of $\mathbf{C}$), i.e., of θ.

It should be emphasized that this reasoning requires that a whitening process precedes the ICA step and that the output signals are normalized to have a unitary variance.

The reasoning held in this paper cannot be easily generalized to all multiple inputs multiple outputs (MIMO) systems. Indeed, the key point in TITO systems is that we have only one degree of freedom for $\mathbf{y}$: the angle θ. In $n \times n$ (with $n > 2$) systems, for a fixed value of an element of a row of $\mathbf{C}$, it remains $n-2$ degrees of freedom to adjust the others of the same row, due to the constraint $E\{\mathbf{y}\mathbf{y}^t\} = \mathbf{I}_n$. Consequently, local maxima for $|\kappa|$ may appear for each value of the fixed coefficient. By contrast, the generalization to two input multiple outputs (TIMO) systems is direct if a principal component analysis is first applied on the mixtures $\mathbf{x}$, to project them on a two-dimensional space, implying that $\mathbf{C}$ remains a 2×2 matrix, and (3) still holds.

Nevertheless, this paper shows that contrast functions for multimodal gradient-based source separation exist that prevent the existence of spurious maxima (at least for $n = 2$), avoiding this well-known drawback when information theoretic contrasts are used.

Acknowledgment

The authors would like to thank Prof. C. Jutten from INP Grenoble, France, for his appreciated contribution to this paper and the anonymous reviewers for their comments, in particular for pointing out that the results of this paper also apply to TIMO systems.

References

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

[2] R. Boscolo, H. Pan, and V. P. Roychowdhury, "Independent component analysis based on nonparametric density estimation," *IEEE Trans. Neural Netw.*, vol. 15, no. 1, pp. 55–65, Jan. 2004.

[3] F. Vrins and M. Verleysen, "On the entropy minimization of a linear mixture of variables for source separation," *Signal Process.*, to be published.

[4] F. Vrins, C. Archambeau, and M. Verleysen, "Entropy minima and distribution structural modifications in blind separation of multi-modal sources," in *Amer. Inst. Phys.*, Jul. 2004, to be published.

[5] E. G. Learned-Miller and J. W. Fisher III, "ICA using spacings estimates of entropy," *J. Mach. Learning Res.*, vol. 4, pp. 1271–1295, 2003.

[6] S. Haykin, Ed., *Unsupervised Adaptive Filtering Vol. 1: Blind Source Separation*. New York: Wiley, 2000, ch. IV, pp. 171–173.

[7] S. Cruces, A. Cichocki, and S. Amari, "The minimum entropy and cumulants based contrast functions for blind source extraction," in *Proc. IWANN, LNCS 2085*, J. Mira and A. Prieto, Eds. New York: Springer-Verlag, 2001, pp. 786–793.

[8] C. E. Shannon, *The Mathematical Theory of Communication*. Chicago, IL: Univ. Illinois Press, 1949.

[9] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Stat.*, vol. 33, pp. 1065–1076, 1962.

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

[11] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[12] A. Mansour and C. Jutten, "What should we say about the kurtosis," *IEEE Signal Process. Lett.*, vol. 6, no. 12, pp. 321–322, Dec. 1999.

[13] J. Friedman, "Exploratory projection pursuit," *J. Amer. Stat. Assoc.*, vol. 82, no. 397, pp. 249–266, 1987.

[14] A. Hyvarinen, "New approximations of differential entropy for independent component analysis and projection pursuit," in *Proc. NIPS*. Cambridge, MA: MIT Press, Oct. 1997, pp. 273–279.

[15] P. Comon, "From source separation to blind equalization, contrast-based approaches," in *Proc. Int. Conf. Image Signal Process.*, Agadir, Morocco, May 3–5.