Mutual Information Approach to Blind Separation of Stationary Sources

Dinh Tuan Pham, *Member, IEEE*

***Abstract*—This paper presents a unified approach to the problem of blind separation of sources, based on the concept of mutual information. This concept is applied to the whole source sequences as stationary processes and thus provides a universal contrast applicable to both the instantaneous and convolutive mixture cases. For practical implementation, we introduce several degraded forms of this contrast, computable from a finite-dimensional distribution of the reconstructed source processes only. From them, we derive several sets of estimating equations, generalizing those considered earlier.**

***Index Terms*—Contrast, convolution, entropy, independent component analysis, Kullback–Leibler divergence, mutual information, separation of sources, stationary process.**

I. Introduction

BLIND separation of sources is a topic which has received much attention recently, as it has many important applications (see, e.g., [2] for a review). Basically, one observes several linear instantaneous or convolutive mixtures of independent signals,[1] called sources, and the problem is to recover them from the observations, *without relying on any specific knowledge of the sources*. In this blind context, a sensible method is to adopt the approach of an independent component analysis (ICA) in which a measure of dependence between the reconstructed sources is minimized. A natural such measure is the mutual information which has been introduced in [3] and implemented in [8]. However, this measure and others which have been proposed rely only on the marginal distribution of the sources, thereby ignoring their temporal properties. The exploitation of such properties can yield better separation in the case of instantaneous mixtures and is crucial in the case of convolutive mixtures. The last case has been less well investigated because of lack of a well-understood methodology: most works in this area adopt an *ad hoc* approach based on canceling cross cumulants. Recently, this author [10] has proposed the use of the mutual information between stationary processes as a contrast function for blind sources separation and discusses some implementation issues. In this work, we further develop this proposal and provide full proofs of results, as the cited paper was only a short version presented at the ICA'99 Workshop. We also provide several sets of estimating equations, obtained by differentiating the above contrast and its variants, which can be related to earlier works [7], [9]. Our emphasis will be on the general ideas and concepts and therefore details of implementations of the methods will not be discussed (these implementations would depend on area of applications and can be the topics of subsequent works).

To proceed, let us describe the problem in mathematical terms and introduce some notations. We assume that K sequences of observations $\{X_k(t), t \in \mathbb{Z}\}$, $k = 1, \ldots, K$, are available, each being a mixture of K independent sources, $\{S_j(t), t \in \mathbb{Z}\}$, $k = 1, \ldots, K$, either instantaneously or through a convolution. More precisely, in the last case

$$\boldsymbol{X}(t) = \sum_{l=-\infty}^{\infty} \boldsymbol{A}(l)\boldsymbol{S}(t-l) = (\boldsymbol{A} \star \boldsymbol{S})(t) \tag{1}$$

where $\boldsymbol{X}(t)$ and $\boldsymbol{S}(t)$ denote the vectors $[X_1(t) \cdots X_K(t)]^{\mathrm{T}}$ and $[S_1(t) \cdots S_K(t)]^{\mathrm{T}}$, respectively, $\{\boldsymbol{A}(l), l \in \mathbb{Z}\}$ is a sequence of matrices, and $\star$ denotes convolution. The instantaneous mixture case can be viewed as a particular case of the above where the sequence $\{\boldsymbol{A}(l) l \in \mathbb{Z}\}$ reduces to a single matrix of index 0, which we denote again by $\boldsymbol{A}$, so that

$$\boldsymbol{X}(t) = \boldsymbol{A}\boldsymbol{S}(t). \tag{2}$$

To separate the sources one naturally performs an "inverse" transformation on the sequence of observed vectors, namely

$$\boldsymbol{Y}(t) = \boldsymbol{B}\boldsymbol{X}(t) \tag{3}$$

in the instantaneous mixture case, $\boldsymbol{B}$ denoting the separation matrix, and

$$\boldsymbol{Y}(t) = \sum_{l=-\infty}^{\infty} \boldsymbol{B}(l)\boldsymbol{X}(t-l) = (\boldsymbol{B} \star \boldsymbol{X})(t) \tag{4}$$

in the convolutive mixture case, $\{\boldsymbol{B}(l), l \in \mathbb{Z}\}$ denoting the sequence of separation matrices.

The idea is to determine the separating matrix $\boldsymbol{B}$ or sequence of matrices $\{\boldsymbol{B}(l), l \in \mathbb{Z}\}$ such that the output sequence $\{\boldsymbol{Y}(t), t \in \mathbb{Z}\}$ in (3) or (4) has components as independent as possible. This is precisely the goal of the ICA, except that ICA thus far has been restricted to instantaneous transformations and the observed sequence $\{\boldsymbol{X}(t), t \in \mathbb{Z}\}$ is not necessarily a mixture of independent sources. Clearly, our approach requires a good measure of dependence between random stationary processes (as we assume that the sources are stationary), which we take to be the mutual information. This is a theoretical measure as it involves the density functions of the source processes, which in practice must be estimated from the data.

Manuscript received July 1, 2000; revised October 15, 2001.

The author is with the Laboratory of Modeling and Computation, IMAG–C.N.R.S., 38041 Grenoble Cedex 9, France (e-mail: Dinh-Tuan.Pham@imag.fr).

Communicated by U. Madhow, Associate Editor for Detection and Estimation.

Publisher Item Identifier S 0018-9448(02)05149-0.

[1]In a more general setup, the output channels may be corrupted with noises, but in this paper we shall restrict ourselves to the pure mixture case.

The estimation problem will not be considered (for a simple case, see [8]) although we are fully aware of the difficulty of estimating the density in a high-dimensional space and we will try to avoid it as much as possible.

For ease of reading, proofs of results will be relegated to the Appendix.

II. Mutual Information Between Stationary Processes

A. Some Definitions and Notations

Recall that the mutual information between a set of K random vectors $Y_1, \ldots, Y_K$, with joint and marginal density functions $p_{Y_1,\ldots,Y_K}$ and $p_{Y_1}, \ldots, p_{Y_K}$, is given by [4]

$$I(Y_1, \ldots, Y_K) = -\mathrm{E}\log\frac{\prod_{i=1}^{K} p_{Y_k}(Y_k)}{p_{Y_1,\ldots,Y_K}(Y_1, \ldots, Y_K)} = \sum_{i=1}^{K} h(Y_k) - h(Y_1, \ldots, Y_K)$$

where

$$h(Y_1, \ldots, Y_K) = -\mathrm{E}\log p_{Y_1,\ldots,Y_K}(Y_1, \ldots, Y_K)$$

and

$$h(Y_k) = -\mathrm{E}\log p_{Y_k}(Y_k)$$

are the (Shannon differential) joint and marginal entropies of $Y_1, \ldots, Y_K$, respectively. Note that the notations

$$h([Y_1^{\mathrm{T}} \ldots Y_K^{\mathrm{T}}]^{\mathrm{T}}) \quad \text{and} \quad h(Y_1, \ldots, Y_K)$$

are the same and will be used interchangeably.

The mutual information is actually the Kullback–Leibler divergence between the joint density of $Y_1, \ldots, Y_K$ and their product densities. From the inequality $\log x \leq x - 1$ for all $x \geq 0$, with equality attained only at $x = 1$, one can see that $I(Y_1, \ldots, Y_K) \geq 0$ with equality if only if $Y_1, \ldots, Y_K$ are independent. Thus, the mutual information is a measure of dependence between a set of random vectors.

1) Entropy of Stationary Processes: The entropy concept can be generalized to the case of stationary (vector) processes. For any process $\{Y(t), t \in \mathbb{Z}\}$ one can write [4]

$$h[Y(1), \ldots, Y(m)] = \sum_{t=1}^{m} h[Y(t)|Y(t-1), \ldots, Y(1)] \quad (5)$$

where

$$h[Y(t)|Y(t-1), \ldots, Y(1)] = h[Y(t), \ldots, Y(1)] - h[Y(t-1), \ldots, Y(1)] \quad (6)$$

is the conditional entropy (that is the expected entropy of the conditional distribution) of $Y(t)$ given $Y(t-1), \ldots, Y(1)$. But the conditioning decreases the entropy (see [4]), hence if the process is stationary

$$h[Y(t+1)|Y(t), \ldots, Y(1)] \leq h[Y(t+1)|Y(t), \ldots, Y(2)] = h[Y(t)|Y(t-1), \ldots, Y(1)] \quad (7)$$

One then deduces the following result, which is somewhat more precise than a result in [4, pp. 64–65 and 273].[2]

Lemma 1: For any stationary process $\{Y(t), t \in \mathbb{Z}\}$

$$h[Y(1), \ldots, Y(m)]/m \geq h[Y(m)|Y(m-1), \ldots, Y(1)]$$

and both sides of this inequality converge nonincreasingly to the same limit (possibly $-\infty$) as $m \to \infty$.

Following [4], we call the common limit in Lemma 1 the entropy (rate) of the process $\{Y(t), t \in \mathbb{Z}\}$ and denote it by $h[Y(\cdot)]$. If $Y(t)$ is a vector with components $Y_1(t), \ldots, Y_K(t)$, it is also called joint entropy (rate) of the processes

$$\{Y_k(t), t \in \mathbb{Z}\}, \qquad k = 1, \ldots, K$$

and denoted by $h[Y_1(\cdot), \ldots, Y_K(\cdot)]$.

2) Mutual Information Between Processes: The mutual information between the K jointly stationary processes $\{Y_k(t), t \in \mathbb{Z}\}$, $k = 1, \ldots, K$ can now be defined as [4]

$$I[Y_1(\cdot), \ldots, Y_K(\cdot)] = \sum_{i=1}^{K} h[Y_k(\cdot)] - h[Y_1(\cdot), \ldots, Y_K(\cdot)]. \quad (8)$$

Clearly, $I[Y_1(\cdot), \ldots, Y_K(\cdot)] \geq 0$ and vanishes if the processes $\{Y_k(t), t \in \mathbb{Z}\}$ are independent. For the converse, we are able to prove it only in the Markovian case (of arbitrary order, however) but we believe it holds much more generally. We can write

$$I[Y_1(\cdot), \ldots, Y_K(\cdot)] = \sum_{i=1}^{K} \{h[Y_k(1)|Y_k(t), t<1] - h[Y_k(1)|\boldsymbol{Y}(t), t<1]\} + \left\{\sum_{i=1}^{K} h[Y_k(1)|\boldsymbol{Y}(t), t<1] - h[\boldsymbol{Y}(1)|\boldsymbol{Y}(t), t<1]\right\} \quad (9)$$

where $h[Y_k(1)|Y_k(t), t < 1]$ stands for

$$\lim_{m\to\infty} h[Y_k(m)|Y_k(m-1), \ldots, Y_k(1)] = \lim_{m\to\infty} h[Y_k(1)|Y_k(0), \ldots, Y_k(2-m)]$$

(the last equality coming from stationarity) and similarly for $h[Y_k(1)|\boldsymbol{Y}(t), t < 1]$ and $h[\boldsymbol{Y}(1)|\boldsymbol{Y}(t), t < 1]$, $\boldsymbol{Y}(t)$ denoting the vector $[Y_1(t) \cdots Y_K(t)]^T$. Since the conditioning decreases the entropy, each term in the sum on the right-hand side of (9) is nonnegative. The last term on this right-hand side represents the mutual information between the components of the conditional distribution of $\boldsymbol{Y}(1)$ given $\boldsymbol{Y}(t)$, $t < 1$ and hence is nonnegative as well. Thus, $I[Y_1(\cdot), \ldots, Y_K(\cdot)] = 0$ implies that all these terms vanish, which entails that i) $Y_k(1)$ is independent of $\{Y_j(t), t < 1, j \neq k\}$, conditionally on $Y_k(t)$, $t < 1$ and ii) $Y_1(1), \ldots, Y_K(1)$ are independent conditionally on $\boldsymbol{Y}(t)$, $t<1$. For Markovian processes, ii) means that the transition probability factors into K factors and i) implies that each factor depends only on $\{Y_k(t), t < 1\}$ for an index k. Since the transition probability of a stationary ergodic Markov process determines its distribution entirely, the processes $\{Y_k(t), t \in \mathbb{Z}\}$ must be independent.

[2]Our result provides further an inequality and the monotonicity of the convergence of its left-hand side.

B. Calculation of Entropy

The computation of the entropy of a process through its definition is not practical as it involves a limiting operation. There are some special cases where this can be avoided, which we now consider.

1) Gaussian Processes: For Gaussian processes, a closed-form formula for the entropy is available. Indeed, let $\{\boldsymbol{Y}(t), t \in \mathbb{Z}\}$ be a Gaussian stationary vector process $\{\boldsymbol{Y}(t), t \in \mathbb{Z}\}$, the conditional distribution of $\boldsymbol{Y}(m)$ given $\boldsymbol{Y}(m-1), \ldots, \boldsymbol{Y}(1)$ is a Gaussian distribution with covariance matrix $\boldsymbol{G}_m$, the error covariance matrix of the best linear predictor of $\boldsymbol{Y}(m)$ based on $\boldsymbol{Y}(m-1), \ldots, \boldsymbol{Y}(1)$, hence by a direct calculation (see also [4])

$$h[\boldsymbol{Y}(m)|\boldsymbol{Y}(m-1), \ldots, \boldsymbol{Y}(1)] = \frac{1}{2}\log\det(2\pi e\boldsymbol{G}_m)$$

where $e = \exp(1)$. Then letting m go to infinity and using the extension of Szegö's theorem to the multivariate case (see, e.g., [6, p. 162]), one gets

$$\begin{aligned} h[\boldsymbol{Y}(\cdot)] &= \frac{1}{2}\log\det(2\pi e\boldsymbol{G}) \\ &= \frac{1}{4\pi}\int_{-\pi}^{\pi}\log\det[4\pi^2 e\boldsymbol{f}(\lambda)]\,d\lambda \end{aligned} \tag{10}$$

where $\boldsymbol{G}$ is the error covariance matrix of the best linear predictor of $\boldsymbol{Y}(t)$ based on $\boldsymbol{Y}(s)$, $s < t$, and $\boldsymbol{f}$ is the spectral density matrix of the process.

2) Temporally Independent and Markovian Processes: For such processes, the following result, which is an easy consequence of the definition of entropy and Lemma 1, is quite useful.

Corollary 1: For a stationary process $\{Y(t), t \in \mathbb{Z}\}$

$$h[Y(\cdot)] \le h[Y(m)|Y(m-1), \ldots, Y(1)]$$

with equality if and only if it is Markovian of order $m-1$, that is, the conditional distribution of $Y(t)$ given $Y(\tau)$, $\tau < t$ depends only on $Y(t-1), \ldots, \ldots, Y(t+1-m)$. In particular, $h[Y(\cdot)] \le h[Y(1)]$ with equality if and only if the process is temporally independent.

Thus, the entropy of a temporally independent process is simply its marginal entropy. Of greater interest is the fact that the entropy of a stationary Markovian process $\{Y(t), t \in \mathbb{Z}\}$ of order m equals

$$\begin{aligned} &h[Y(m)|Y(m-1), \ldots, Y(1)] \\ &\qquad = h[Y(1), \ldots, Y(m)] - h[Y(1), \ldots, Y(m-1)] \end{aligned}$$

Because of stationarity, the last right-hand side also equals

$$\begin{aligned} &h[Y(1), \ldots, Y(m)] - h[Y(2), \ldots, Y(m)] \\ &\qquad = h[Y(1)|Y(2), \ldots, Y(m)]. \end{aligned}$$

3) Filtered Processes: A general class of processes which includes the widely used autoregressive moving average (ARMA) processes is the class of *linear processes*, defined as follows.

Definition 1: A process $\{\boldsymbol{Y}(t), t \in \mathbb{Z}\}$ is called linear if it can be represented in the form

$$\boldsymbol{Y}(t) = \sum_{l=-\infty}^{\infty}\boldsymbol{A}(l)\boldsymbol{e}(t-l) \tag{11}$$

where $\{\boldsymbol{A}(l), l \in \mathbb{Z}\}$ is some sequence of matrices and $\{\boldsymbol{e}(t), t \in \mathbb{Z}\}$ is a sequence of independent and identically distributed random vectors.

In other words, a linear process is the output of some filter applied to a temporally independent process. Note that we allow the filter to be noncausal ($\boldsymbol{A}(l) \neq \boldsymbol{0}$ for $l < 0$), the causal case will receive some special attention later. Since we already know the entropy of a temporally independent process, it would be a simple matter to compute that of a linear process if we *knew* how to relate the entropy of a filtered process to that of the original process. To derive such a result, we will need to restrict ourselves to a class of "well-behaved" filters. We call *a sequence of square matrices $\{\boldsymbol{B}(l),\ l \in \mathbb{Z}\}$ of class $\mathcal{A}$ if*

$$\sum_{l=-\infty}^{\infty}\|\boldsymbol{B}(l)\| < \infty$$

and

$$\det\left[\sum_{l=-\infty}^{\infty}\boldsymbol{B}(l)e^{jl\lambda}\right] \neq 0$$

for all λ. It can then be seen that for such a sequence, the process $\{\boldsymbol{Y}(t), t \in \mathbb{Z}\}$ defined in (4) is well defined for any stationary process $\{\boldsymbol{X}(t), t \in \mathbb{Z}\}$ with finite αth absolute moment ($\alpha \ge 1$) and is itself stationary with finite αth absolute moment. Further, the class $\mathcal{A}$ is closed with respect to the convolution in the sense that if it contains the sequences $\{\boldsymbol{B}(l), l \in \mathbb{Z}\}$ and $\{\boldsymbol{C}(l), l \in \mathbb{Z}\}$, then it also contains their convolution.[3] Another interesting property of the class $\mathcal{A}$ is that any sequence $\{\boldsymbol{B}(l), l \in \mathbb{Z}\}$ in this class admits an inverse, with respect to the convolution, which is also of this class, the inverse being precisely the sequence of the Fourier coefficients of the function

$$\lambda \mapsto \left[\sum_{l=-\infty}^{\infty}\boldsymbol{B}(l)e^{il\lambda}\right]^{-1}.$$

This result follows from a result of Wiener which says that if f is a 2π-periodic function, nonzero everywhere, and has absolutely summable Fourier coefficients, then the same is true for $1/f$ (see, e.g., [14, p. 245]).

We shall further need a lower semi-continuity condition of the entropy functional.

Definition 2: The entropy functional is said to be lower semi-continuous, with respect to the convolution, at the process $\{\boldsymbol{Y}(t), t \in \mathbb{Z}\}$ if for any integer $m > 0$, real $\epsilon > 0$, there exists $\delta > 0$ such that

$$h[(\boldsymbol{C} \star \boldsymbol{Y})(1), \ldots, (\boldsymbol{C} \star \boldsymbol{Y})(m)] \le h[\boldsymbol{Y}(1), \ldots, \boldsymbol{Y}(m)] + \epsilon$$

for all sequences $\{\boldsymbol{C}(l), l \in \mathbb{Z}\}$ satisfying

$$\|\boldsymbol{C}(0) - \boldsymbol{I}\| + \sum_{l \neq 0}\|\boldsymbol{C}(l)\| \le \delta.$$

[3] This can be seen from

$$\begin{aligned} \sum_l\left\|\sum_m \boldsymbol{B}(l-m)\boldsymbol{C}(m)\right\| &\le \sum_l\sum_m \|\boldsymbol{B}(l-m)\|\,\|\boldsymbol{C}(m)\| \\ &= \left[\sum_l\|\boldsymbol{B}(l)\|\right]\left[\sum_m\|\boldsymbol{C}(m)\|\right] \end{aligned}$$

and the fact that the Fourier transform transforms a convolution into a multiplication.

Proposition 1: Let $\{\boldsymbol{X}(t), t \in \mathbb{Z}\}$ be a vector random stationary process admitting αth absolute moment ($\alpha \geq 1$) and $\boldsymbol{Y}(t) = (\boldsymbol{B} \star \boldsymbol{X})(t)$, where $\{\boldsymbol{B}(l), l \in \mathbb{Z}\}$ is a sequence of matrices of class $\mathcal{A}$. Assume that the entropy functional is lower semi-continuous at both $\{\boldsymbol{X}(t), t \in \mathbb{Z}\}$ and $\{\boldsymbol{Y}(t), t \in \mathbb{Z}\}$, then

$$h[\boldsymbol{Y}(\cdot)] = h[\boldsymbol{X}(\cdot)] + \int_{-\pi}^{\pi} \log \left| \det \sum_{j=-\infty}^{\infty} \boldsymbol{B}(l) e^{il\lambda} \right| \frac{d\lambda}{2\pi}.$$

Proposition 1 can be viewed as an extension of the following result, which can be easily obtained along the same lines as in [8] based on [8, Lemma A1]: *Let $\boldsymbol{X}$ be a random vector and $\boldsymbol{B}$ be an invertible matrix, then the entropy of $\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{X}$ equals*

$$h(\boldsymbol{Y}) = h(\boldsymbol{X}) + \log |\det \boldsymbol{B}|. \tag{12}$$

The lower semi-continuity condition is admittedly hard to verify, but it is very mild. It holds under the following condition, which we believe to be far from necessary.

Lemma 2: Let $\{\boldsymbol{Y}(t), t \in \mathbb{Z}\}$ be a vector random stationary process admitting the αth absolute moment ($\alpha \geq 1$) such that the joint density p_m of $\boldsymbol{Y}(1), \ldots, \boldsymbol{Y}(m)$ exists (for all m) and is differentiable with $\|\nabla \log p_m(\boldsymbol{y})\| \leq C(1 + \|\boldsymbol{y}\|^{\alpha-1})$, ∇ denoting the gradient operator and C being a constant.[4] Then the entropy functional is lower semi-continuous at $\{\boldsymbol{Y}(t), t \in \mathbb{Z}\}$.

Proposition 1 is fundamental in that it describes how the entropy changes when the process is filtered. In particular, it provides the entropy of a linear process or more generally a filtered Markov process. Further, it provides a method for the *deconvolution* of a linear (or a filtered Markov) process, through the use of the following corollary.

Corollary 2: Let $\{\boldsymbol{Y}(t), t \in \mathbb{Z}\}$ be a vector stationary process such that the entropy functional is lower semi-continuous at $\{(\boldsymbol{B} \star \boldsymbol{Y})(t), t \in \mathbb{Z}\}$, for all sequences $\{\boldsymbol{B}(l), l \in \mathbb{Z}\}$ of class $\mathcal{A}$. Then

$$h[\boldsymbol{Y}(\cdot)] \leq h[(\boldsymbol{B} \star \boldsymbol{Y})(1)] - \int_{-\pi}^{\pi} \log \left| \det \sum_{j=-\infty}^{\infty} \boldsymbol{B}(l) e^{il\lambda} \right| \frac{d\lambda}{2\pi} \tag{13}$$

with equality if and only if the process $\{(\boldsymbol{B} \star \boldsymbol{Y})(t), t \in \mathbb{Z}\}$ is temporally independent. The same inequality holds with $h[(\boldsymbol{B} \star \boldsymbol{Y})(1)]$ replaced by

$$h[(\boldsymbol{B} \star \boldsymbol{Y})(m) | (\boldsymbol{B} \star \boldsymbol{Y})(m-1), \ldots, (\boldsymbol{B} \star \boldsymbol{Y})(1)];$$

in this case, equality is attained if and only if the process $\{(\boldsymbol{B} \star \boldsymbol{Y})(t), t \in \mathbb{Z}\}$ is Markovian of order $m-1$.

Clearly, Corollary 2 still holds if the class $\mathcal{A}$ is replaced by a smaller subclass. A subclass of interest is the subclass $\mathcal{A}^+$ of sequences $\{\boldsymbol{B}(l), l \in \mathbb{Z}\}$ which are causal and have causal inverses, in the sense that it and its inverse sequence vanish at negative indexes. It is well known that the last condition is equivalent to the minimum-phase condition: $\det[\sum_{l=0}^{\infty} \boldsymbol{B}(l) z^l] \neq 0$ for all complex number z of modulus not exceeding 1. Under this condition

$$\oint_C \left\{ \log \det \left[\boldsymbol{I} + \sum_{l=1}^{\infty} \boldsymbol{B}(0)^{-1} \boldsymbol{B}(l) z^l \right] \right\} dz/(2\pi i z) = 0$$

[4]This constant can depend on m as the δ in Definition 2 can depend on m.

the integration being made along the unit circle C of the complex plane. Thus,

$$\int_{-\pi}^{\pi} \log \left| \det \sum_{l=0}^{\infty} \boldsymbol{B}(l) e^{il\lambda} \right| \frac{d\lambda}{2\pi} = \log \det |\boldsymbol{B}(0)|. \tag{14}$$

Therefore, by restricting to the class $\mathcal{A}^+$, one gets the same result as in Corollary 2 with the last integral in (13) replaced by $\log \det |\boldsymbol{B}(0)|$.

III. Contrasts

We shall assume throughout that the sequence $\{\boldsymbol{A}(l), l \in \mathbb{Z}\}$ in (1) is of class $\mathcal{A}$, hence in the reconstruction formula (4) we will restrict ourselves to sequences $\{\boldsymbol{B}(l), l \in \mathbb{Z}\}$ of this class. As mentioned earlier, to separate the source, one may minimize the mutual information $I[Y_1(\cdot), \ldots, Y_K(\cdot)]$, where $Y_k(t)$ are the components of $\boldsymbol{Y}(t)$ defined by (3) or (4) according to model (2) or (1) is considered. But by Proposition 1, this criterion equals, up to a constant term

$$C_\infty = \sum_{k=1}^{K} h[Y_k(\cdot)] - \int_{-\pi}^{\pi} \log \left| \det \sum_{l=-\infty}^{\infty} \boldsymbol{B}(l) e^{il\lambda} \right| \frac{d\lambda}{2\pi}. \tag{15}$$

in the case of model (1), or the same expression but with the integral replaced by $\log |\det \boldsymbol{B}|$, in the case of model (3).

The above criterion, by construction, is a *contrast* [3] in the sense that it is minimized if the reconstructed sources $\{Y_k(t), t \in \mathbb{Z}\}$, $k = 1, \ldots, K$, coincide with the true sources *up to a permutation and a filtering* (or a scaling in the instantaneous mixture case). This ambiguity is *inherent* to the blind source separation problem (since it relies only on the independence assumption of the sources) and is manifested in the invariance property of C_∞: it is unchanged when one preconvolves the sequence $\{\boldsymbol{B}(l), l \in \mathbb{Z}\}$ with a sequence of diagonal matrices of class $\mathcal{A}$ and premultiplies the result with a permutation matrix, as can be easily seen by applying Proposition 1 to scalar filtered processes. To be useful, however, C_∞ should be discriminating, in the sense of [3], that is, it should attain its minimum *only* when the reconstructed sources coincide with the true sources *modulo the above ambiguity*. But minimizing C_∞ can only ensure the independence between the reconstructed sources and thus for this contrast to be discriminating one may need some further conditions (such as non-Gaussianity), but since we will not actually use it, we do not pursue this question.

A. Contrasts for Instantaneous Mixtures

The contrast C_∞ is of theoretical interest only, since its computation requires the complete knowledge of the distribution of each process $\{Y_k(t), t \in \mathbb{Z}\}$. Although one can always approximate $h[Y_k(\cdot)]$ by $h[Y_k(1), \ldots, Y_k(m)]/m$, the number m might be very large for the approximation to be accurate, leading to the problem of estimation of a density in a *high-dimensional* space, which we would like to avoid.[5] Therefore, it is of interest to obtain a simplified version of C_∞.

[5]The amount of data needed for a "good" density estimation in a high-dimensional space grows *exponentially* with the dimension.

1) Contrasts Based on Finite Joint Distribution: Instead of considering the mutual information between processes, we consider the mutual information between segments of processes. Explicitly, we consider the criterion

$$\frac{1}{m} I\{[Y_1(1)\cdots Y_1(m)]^{\mathrm{T}}, \ldots, [Y_K(1)\cdots Y_K(m)]^{\mathrm{T}}\} \quad (16)$$

where m is a (small) integer and $Y_k(t)$ are the components of $\boldsymbol{Y}(t)$, defined in (3). But from (12)

$$h[\boldsymbol{Y}(1), \ldots, \boldsymbol{Y}(m)] = m \log\det(\boldsymbol{B}) + h[\boldsymbol{X}(1), \ldots, \boldsymbol{X}(m)] \quad (17)$$

hence this criterion can be seen to be equal, up to a constant term, to

$$C_m(\boldsymbol{B}) = \frac{1}{m}\sum_{k=1}^{K} h[Y_k(1), \ldots, Y_k(m)] - \log|\det \boldsymbol{B}|. \quad (18)$$

By construction, this is a contrast, which, in the case where $m = 1$, has been shown to be discriminating if no more than one source can be Gaussian [3]. For $m > 1$, one can allow the sources to be Gaussian if the covariance matrices of m consecutive observations of the Gaussian sources are not proportional (see [12], [13]).

One can view C_m as an approximation to C_∞, in which the entropy $h[Y_k(\cdot)]$ is replaced by $h[Y_k(1), \ldots, Y_k(m)]/m$. An alternative approach is to replace it by the conditional entropy $h[Y_k(m)|Y_k(m-1), \ldots, Y_k(1)]$, which is a better approximation by Lemma 1. This leads to the criterion

$$C_m^*(\boldsymbol{B}) = \sum_{k=1}^{K} h[Y_k(m)|Y_k(m-1), \ldots, Y_k(1)] - \log|\det \boldsymbol{B}|. \quad (19)$$

The following result shows that $C_m^*(\boldsymbol{B})$ is a contrast. It can be shown to be discriminating under the same conditions as for the contrast C_m (see [12], [13]).

Lemma 3: Under the model (2), the criterion (19) equals the expected Kullback–Leibler divergence between the conditional distribution of $\boldsymbol{Y}(m)$ given $\boldsymbol{Y}(m-1), \ldots, \boldsymbol{Y}(1)$ and the product of the conditional distribution of $Y_k(m)$ given $Y_k(m-1), \ldots, Y_k(1)$, plus the constant term $h[\boldsymbol{X}(m)|\boldsymbol{X}(m-1), \ldots, \boldsymbol{X}(1)]$.

2) The Gaussian Mutual Information: To avoid the difficulty in calculating the entropy and mutual information, we introduce the Gaussian entropy mutual information, defined as before but with the random vectors or processes involved replaced by the Gaussian random vectors or processes having the same covariance structure. From (8) and (10), the Gaussian mutual information between the stationary processes $\{Y_k(t), t \in \mathbb{Z}\}$, $k = 1, \ldots, K$ is

$$I_g[Y_1(\cdot), \ldots, Y_K(\cdot)] = \frac{1}{4\pi}\int_{-\pi}^{\pi} [\log\det \operatorname{diag} \boldsymbol{f}_{\boldsymbol{Y}}(\lambda) - \log\det \boldsymbol{f}_{\boldsymbol{Y}}(\lambda)]\, d\lambda \quad (20)$$

where $\boldsymbol{f}_{\boldsymbol{Y}}$ is the spectral density matrix of the vector process $\{\boldsymbol{Y}(t) = [Y_1(t)\cdots Y_K(t)]^{\mathrm{T}}, t \in \mathbb{Z}\}$ and diag denotes the diagonal matrix with the same diagonal as its argument.

The criterion (20) is a joint diagonalization criterion, since it is nonnegative and can be zero if and only if $\boldsymbol{f}_{\boldsymbol{Y}}$ is diagonal almost everywhere. This is easily seen from the Hadamard inequality which says that for a positive-definite matrix $\boldsymbol{f}$, $\det \boldsymbol{f} < \det \operatorname{diag} \boldsymbol{f}$ unless $\boldsymbol{f}$ is diagonal in which case one has equality (see, e.g., [4, p. 502]). Since this criterion involves only the correlations between the sources, it would *not permit the separation of a convolutive mixture*: It is easy to see that it vanishes as soon as $[\sum_{l=-\infty}^{\infty} \boldsymbol{B}(l)e^{il\lambda}]\boldsymbol{f}_{\boldsymbol{Y}}^{1/2}(\lambda)$ is a unitary matrix for almost all λ, $\boldsymbol{f}_{\boldsymbol{Y}}^{1/2}$ being the Cholesky factor in the decomposition $\boldsymbol{f}_{\boldsymbol{Y}} = \boldsymbol{f}_{\boldsymbol{Y}}^{1/2}(\boldsymbol{f}_{\boldsymbol{Y}}^{1/2})^{\mathrm{T}}$. However, in the instantaneous mixture case, this contrast is discriminating, provided that there exists no pair of sources which have proportional spectral densities [11]. Since $\log\det \boldsymbol{f}_{\boldsymbol{Y}} = 2\log\det \boldsymbol{B} + \log\det \boldsymbol{f}_{\boldsymbol{X}}$ in this case, this contrast is equivalent to

$$C_g(\boldsymbol{B}) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \log[\det \operatorname{diag} \boldsymbol{f}_{\boldsymbol{Y}}(\lambda)]\, d\lambda - \log\det \boldsymbol{B}.$$

One can further degrade the above contrast by considering the Gaussian analogs of the criteria C_m and C_m^*. This yields, after dropping a constant term

$$C_{g,m}(\boldsymbol{B}) = \frac{1}{2m}\sum_{k=1}^{K} \log\left|\det \operatorname{cov}\left\{[Y_k(1)\cdots Y_k(m)]^{\mathrm{T}}\right\}\right| - \log|\det \boldsymbol{B}| \quad (21)$$

$$C_{g,m}^*(\boldsymbol{B}) = \frac{1}{2}\sum_{k=1}^{K} \log \operatorname{var}\{Y_k(m) - Y_k(m|1:m-1)\} - \log|\det \boldsymbol{B}| \quad (22)$$

where $\operatorname{cov}\{\cdot\}$ refers to covariance matrix, $\operatorname{var}\{\cdot\}$ refers to variance, and $Y_k(m|1 : m-1)$ denotes the best linear predictor of $Y_k(m)$ based on $Y_k(1), \ldots, Y_k(m-1)$. It can be shown that (21) and (22) are discriminating contrasts provided that there exists no pair j, k such that the covariance matrices of $[S_j(1)\cdots S_j(m)]^{\mathrm{T}}$ and $[S_k(1)\cdots S_k(m)]^{\mathrm{T}}$ are proportional [11].

B. Convolutive Mixtures and/or Linear Source Processes

Unfortunately, the above approach cannot be generalized to the case of convolutive mixtures. The reason is that the convolution is a transformation on the whole process, not a finite segment of it. The criterion (16) is still a contrast (although we are not sure if it is discriminating), but, unlike the instantaneous mixture case, would involve the joint entropy of $\boldsymbol{Y}(1), \ldots, \boldsymbol{Y}(m)$; Proposition 1 is not applicable since it concerns the entropy of a whole process, not of a finite segment of it. Thus, the use of (16) would require the estimation of the entropy of an mK-dimensional distribution, which we would like to avoid. By the same reason, there is no analog of the contrast (19) for the convolutive mixture case; Lemma 3 applies only to the instantaneous mixture case.

However, if one restricts oneself to the class of linear or Markovian source processes, then simple contrasts can be constructed.

1) Convolutive Mixtures of Linear Sources:

Proposition 2: Assume that the sources are linear processes, specifically

$$S_k(t) = \sum_{l=-\infty}^{\infty} a_k(l) e_k(t-l) \tag{23}$$

where $\{e_k(t),\, t \in \mathbb{Z}\}$ are temporally independent processes and $\{a_k(l),\, l \in \mathbb{Z}\}$ are sequences of class $\mathcal{A}$, then the criterion

$$C_1[\boldsymbol{B}(\cdot)] = \sum_{k=1}^{K} h[Y_k(1)] - \int_{-\pi}^{\pi} \log \left| \det \sum_{l=-\infty}^{\infty} \boldsymbol{B}(l) e^{il\lambda} \right| \frac{d\lambda}{2\pi}. \tag{24}$$

is minimized if and only if the processes $\{Y_k(t),\, t \in \mathbb{Z}\}$ are independent among themselves and are temporally independent.

The preceding result shows that the criterion (24) is a contrast since it is minimized when the reconstructed sources coincide with the true sources up to a permutation and a convolution. However, it is not necessarily discriminating since there is still the possibility that the processes $\{Y_k(t),\, t \in \mathbb{Z}\}$, despite being independent, do not coincide with the sources up to a permutation and a filtering (an example is the case where the sources are Gaussian). Nevertheless, it can be proved by a different method in [12] (and mentioned in [13]) that the contrast (24) is discriminating if no more than one source can be Gaussian.

As made clear by Proposition 2, minimizing (24) not only separates the sources but deconvolves them as well. If there can be no more than one Gaussian source, one would recover the sequences $\{e_k(t),\, t \in \mathbb{Z}\}$ in (23) up to a scaling, a permutation, and a time shift. More precisely, minimizing (24) yields $Y_k(t) = \alpha_k e_{\pi_k}(t - \tau_k)$ for some permutation $\{\pi_1, \ldots, \pi_K\}$, some nonzero constants $\alpha_1, \ldots, \alpha_k$, and some integers $\tau_1, \ldots, \tau_k$. Note that, unlike the contrast (15) which is invariant with respect to filtering, the contrast (24) *is not*.

A more restrictive assumption on the distribution of the sources is that they are *linear causal processes with minimum phase*. By this we mean that the sequences $\{a_k(l),\, l \in \mathbb{Z}\}$ in the representation (23) are of class $\mathcal{A}^+$. Assume further that the sequence $\{\boldsymbol{A}(l),\, l \in \mathbb{Z}\}$ in (1) is also of this class; then it makes sense to restrict the sequence $\{\boldsymbol{B}(l),\, l \in \mathbb{Z}\}$ in (4) to this class as well. Therefore, the contrast (24), by (14), reduces to

$$C_1^+[\boldsymbol{B}(\cdot)] = \sum_{k=1}^{K} h[Y_k(1)] - \log |\det \boldsymbol{B}(0)|. \tag{25}$$

As before, minimizing it *among all sequences of class* $\mathcal{A}^+$ not only separates the sources but deconvolves them as well. The contrast (25), however, relies on somewhat artificial assumptions on the sources and the mixing matrix sequence. But it has the advantage of being simple.

2) Convolutive Mixtures of Markovian Sources: A weaker assumption on the sources is that they are filtered Markov processes. More precisely, it is assumed that the sources can be represented by (23) but with $\{e_k(t),\, t \in \mathbb{Z}\}$ now being an $(m-1)$th-order Markov process. Then, similarly to Proposition 2, one can show that the criterion

$$C_m^*[\boldsymbol{B}(\cdot)] = \sum_{k=1}^{K} h[Y_k(m) | Y_k(m-1), \ldots, Y_k(1)] - \int_{-\pi}^{\pi} \log \left| \det \sum_{l=-\infty}^{\infty} \boldsymbol{B}(l) e^{il\lambda} \right| \frac{d\lambda}{2\pi} \tag{26}$$

is minimized if and only if the processes $\{Y_k(t), t \in \mathbb{Z}\}$ are independent among themselves and are Markovian of order m–1.

The proof of this result is very similar to that of Proposition 2, substituting $h[Y_k(1)]$ by $h[Y_k(m)|Y_k(m-1), \ldots, Y_k(1)]$.

As before, minimizing $C_m^*[\boldsymbol{B}(\cdot)]$ not only separates the sources but actually extracts the underlying Markov processes which generate them. In practice, it is likely that the sources are themselves Markovian and not filtered Markov processes; in this case, they are recovered exactly up to a permutation, a scaling, and a time shift. The ambiguity with respect to filtering is lifted because one has the *a priori* information that the sources are Markovian.

3) Instantaneous Mixture of Linear Sources: In the previous subsection, we have focused on the convolutive mixture case, but the approach there can be also applied to the instantaneous mixture case. By Corollary 2, C_∞ in this case is bounded above by

$$\sum_{k=1}^{K} \inf \left\{ h[(b_k \star Y_k)(1)] - \int_{-\pi}^{\pi} \log \left| \sum_{l=-\infty}^{\infty} b_k(l) e^{il\lambda} \right| \frac{d\lambda}{2\pi} \right\} - \log |\det \boldsymbol{B}| \tag{27}$$

where the infimum is taken over all sequences $\{b_k(l),\, l \in \mathbb{Z}\}$ of class $\mathcal{A}$. Further, equality can be achieved if and only if the process $\{Y_k(t),\, t \in \mathbb{Z}\}$ is linear. Since C_∞ is a contrast and the sources are linear processes, this shows that (27) is indeed a contrast.

Clearly, by (14), C_∞ is also bounded above by

$$\sum_{k=1}^{K} \inf \{ h[(b_k \star Y_k)(1)] - \log |b_k(0)| \} - \log |\det \boldsymbol{B}| \tag{28}$$

where the infimum is taken over all sequences $\{b_k(l),\, l \in \mathbb{Z}\}$ of class $\mathcal{A}^+$. Further, by Corollary 2 again, equality can be achieved if and only if the processes $\{Y_k(t),\, t \in \mathbb{Z}\}$ are linear causal with minimum phase. Thus, in the case where the sources are linear causal with minimum phase processes, (28) is indeed a contrast.

It is worthwhile to note that the coefficients $b_k(0)$ in (27) and (28) can be taken equal to 1, that is, the infima there are taken over all sequences $\{b_k(l),\, l \in \mathbb{Z}\}$ in $\mathcal{A}$ or $\mathcal{A}^+$ with $b_k(0) = 1$. This is because multiplying the sequence $\{b_k(l),\, l \in \mathbb{Z}\}$ by a constant does not change the expression inside the curly bracket $\{\,\}$ in (27) and (28).

C. Discussion

The contrast C_1 is well known (see, e.g., [3], [8]). But it exploits only the marginal distribution of the sources at a given time point. Our contrasts C_m and C_m^*, $m > 1$, involve their temporal dependence as well and thus could have better per-

formance especially in the case where the sources are strongly temporally dependent (note that if they are white, $C_m = C_m^* = C_1$). However, m should be small due to the difficulty of estimating the entropy of a high-dimensional random vector. Thus, it might be of interest to consider the contrasts (27) and (28) which requires only the entropy of random variables and yet taking into account the temporal dependence of the sources. The drawback is that they rely on the linearity assumption of the sources and require an extra minimization. Another possibility is to focus only on the second-order dependence of the sources, as implied by the use of the contrasts $C_{m,g}$, $C_{m,g}^*$, and C_g. The use of correlations only, but including lagged correlations, for blind sources separation, has been proposed, for example, in [1] and [9]. The use of C_1 and C_m^* in the convolutive mixtures case is new.

IV. Estimating Equations

By differentiating the above contrasts, one obtains a system of equations to be satisfied, called estimating equations (see [5]). For this purpose, the following result plays a central role.

Lemma 4: Let $\boldsymbol{Y}$ and $\boldsymbol{Z}$ be two random vectors admitting absolute αth moment for some $\alpha \geq 1$. Assume that $\boldsymbol{Y}$ and $\boldsymbol{Y} + \mathcal{E}\boldsymbol{Z}$, $\mathcal{E}$ being a matrix for which the product $\mathcal{E}\boldsymbol{Z}$ makes sense and has the same dimension as $\boldsymbol{Y}$, admit densities $p_{\boldsymbol{Y}}$ and $p_{\boldsymbol{Y}+\mathcal{E}\boldsymbol{Z}}$ satisfying the following conditions.

C1) As $\mathcal{E} \to \mathbf{0}$

$$\int \log[p_{\boldsymbol{Y}}(u)/p_{\boldsymbol{Y}+\mathcal{E}\boldsymbol{Z}}(u)]\, p_{\boldsymbol{Y}+\mathcal{E}\boldsymbol{Z}}(u)\, du \to 0$$

faster than $||\mathcal{E}||$.

C2) The function $-\log p_{\boldsymbol{Y}}$ admits almost everywhere a gradient (column) vector $\psi_{\boldsymbol{Y}}$ such that

$$||\psi_{\boldsymbol{Y}}(u)|| \leq C(1 + ||u||^{\alpha-1}), \qquad \text{for all } u$$

for some constant C.

Then as $\mathcal{E} \to \mathbf{0}$

$$h(\boldsymbol{Y} + \mathcal{E}\boldsymbol{Z}) - h(\boldsymbol{Y}) = \mathrm{E}(\psi_{\boldsymbol{Y}}^T \mathcal{E}\boldsymbol{Z}) + o(\mathcal{E})$$

where $o(\mathcal{E})$ denotes a term tending to $\mathbf{0}$ faster than $\mathcal{E}$.

Note: Condition C1) could be hard to verify, but it is quite reasonable. Indeed

$$\int \log \frac{p_{\boldsymbol{Y}}(u)}{p_{\boldsymbol{Y}+\mathcal{E}\boldsymbol{Z}}(u)}\, p_{\boldsymbol{Y}+\mathcal{E}\boldsymbol{Z}}(u)\, du = \int \left[\log \frac{p_{\boldsymbol{Y}}(u)}{p_{\boldsymbol{Y}+\mathcal{E}\boldsymbol{Z}}(u)} - \frac{p_{\boldsymbol{Y}}(u)}{p_{\boldsymbol{Y}+\mathcal{E}\boldsymbol{Z}}(u)} + 1\right] p_{\boldsymbol{Y}+\mathcal{E}\boldsymbol{Z}}(u)\, du.$$

For small $\mathcal{E}$, one would expect that the expression inside the bracket [ ] is of the order $||\mathcal{E}||^2$ and thus the whole integral would be of this order. The difficulty is that $p_{\boldsymbol{Y}}$ and $p_{\boldsymbol{Y}+\mathcal{E}\boldsymbol{Z}}$ converge to zero at infinity and hence the behavior of the ratio $p_{\boldsymbol{Y}}/p_{\boldsymbol{Y}+\mathcal{E}\boldsymbol{Z}}$ near infinity is difficult to predict. The expression inside the bracket is in general of the order $||\mathcal{E}||^2$ for fixed u, but not uniformly in u. This uniformity is, however, not at all necessary since we will integrate with respect to $p_{\boldsymbol{Y}+\mathcal{E}\boldsymbol{Z}}$, which can be expected to converge to zero with a fast rate. But we have been unable to find simple conditions to ensure that C1) is satisfied.

The function $\psi_{\boldsymbol{Y}}$ will play a fundamental role in the sequel. In the case of a real random variable, it is usually referred to (in the statistical literature) as the score function. For a random vector, we therefore call $\psi_{\boldsymbol{Y}}$ the multivariate score function of the density of $\boldsymbol{Y}$.

A. Instantaneous Mixtures

We now apply the above result to obtain necessary conditions for the contrasts (18) and (19) to be minimized.

Proposition 3: A necessary condition for C_m to be minimized at $\boldsymbol{B}$ is

$$\mathrm{E}\{[Y_j(1)\cdots Y_j(m)]\psi_{k,m}[Y_k(1), \ldots, Y_k(m)]\} = 0, \qquad 1 \leq j \neq k \leq K \quad (29)$$

and for C_m^* to be minimized at $\boldsymbol{B}$ is

$$\mathrm{E}\{[Y_j(1)\cdots Y_j(m)]\psi_{k,m}^*[Y_k(1), \ldots, Y_k(m)]\} = 0, \qquad 1 \leq j \neq k \leq K \quad (30)$$

where $\psi_{k,m}$ and $\psi_{k,m}^*$ are the multivarate score functions of the joint density of $Y_k(1), \ldots, Y_k(m)$ and of the conditional density of $Y_k(m)$ given $Y_k(m-1), \ldots, Y_k(1)$. (Here the conditional density is considered as a function of both the dependent and the conditioning variables.)

Note: It can be seen from the proof of the preceding result that the conditions

$$\mathrm{E}\{[Y_k(1)\cdots Y_k(m)]\psi_{k,m}[Y_k(1), \ldots, Y_k(m)]\} = m$$
$$\mathrm{E}\{[Y_k(1)\cdots Y_k(m)]\psi_{k,m}^*[Y_k(1), \ldots, Y_k(m)]\} = 1$$

are also necessary. But these conditions are actually always satisfied because of the definitions of $\psi_{k,m}$ and $\psi_{k,m}^*$. This is an easy consequence of the following result which can be obtained through an integration by parts.

Lemma 5: Let Y be a random vector having a density f_Y such that $f_Y(y)y \to 0$ as $y \to \pm\infty$ and $-\log f_Y$ admits a gradient ψ_Y. Then $\mathrm{E}[Y_k\psi_{k,Y}(Y)] = 1$, Y_k, and $\psi_{k,Y}$ denotes the kth component of Y and ψ_Y.

Consider now the Gaussian contrasts (21), (22), and (20).

Proposition 4: A necessary condition for $C_{g,m}$ to be minimized at $\boldsymbol{B}$ is

$$\sum_{t=1}^{m} \mathrm{cov}\left\{Y_j(t), \sum_{s=1}^{m} b_{k,m}(t,s)Y_k(s)\right\} = 0, \qquad 1 \leq j \neq k \leq K \quad (31)$$

for $C_{g,m}^*$ to be minimized at $\boldsymbol{B}$ is

$$\sum_{t=1}^{m} \mathrm{cov}\left\{Y_j(t), a_{k,m-1}(m-t)\sum_{l=0}^{m-1} a_{k,m-1}(l)Y_k(m-l)\right\} = 0, \qquad 1 \leq j \neq k \leq K \quad (32)$$

and for C_g^* to be minimized at $\boldsymbol{B}$ is

$$\int_{-\pi}^{\pi} [f_{Y_kY_j}(\lambda)/f_{Y_kY_k}(\lambda)]d\lambda = 0, \quad 1 \leq j \neq k \leq K \quad (33)$$

where $b_{k,m}(t, s)$ are the general elements of the inverse of $\mathrm{cov}\{[Y_k(1)\cdots Y_k(m)]^{\mathrm{T}}\}$, $a_{k,m-1}(l)$ are the coefficients in the representation $\sum_{l=0}^{m-1} a_{k,m-1}(l)Y_k(m-l)$ of $Y_k(m) - Y_k(m|1 : m-1)$ and $f_{Y_kY_j}$ is the cross-spectral density between the processes $\{Y_k(t), t\in\mathbb{Z}\}$ and $\{Y_j(t), t\in\mathbb{Z}\}$.

One can see that (33) is a limiting form of (32) as $m \to \infty$. Indeed, $a_{k,m-1}(l)$ converges to $a_k(l)$ such that $\sum_{l=0}^{\infty} a_k(l)Y_k(t-l)$ is the error of the best linear predictor of $Y_k(t)$ based on $Y_k(s)$, $s < t$. Hence, the right-hand side of (32) converges to a constant times that of (33), since $f_{Y_kY_k}(\lambda)$ is proportional to $1/|\sum_{l=0}^{\infty} a_k(l)e^{il\lambda}|^2$. It is also possible to prove that (33) is a limiting form of (31) as well.

B. Linear and Markovian Sources

We first consider the convolutive mixture case.

Proposition 5: Assume that the sources are linear processes. A necessary condition for the contrast (24) to be minimized at the sequence $\{\boldsymbol{B}(l),\ l \in \mathbb{Z}\}$ of class $\mathcal{A}$ is

$$\mathrm{E}\{Y_j(1-\tau)\psi_k[Y_k(1)]\} = 0, \qquad j,\ k = 1, \ldots, K,\ \tau \in \mathbb{Z},\ j \neq k \text{ or } \tau \neq 0 \quad (34)$$

and if the sources are also causal with minimum phase, a necessary condition for the contrast (25) to be minimized at a sequence $\{\boldsymbol{B}(l),\ l \in \mathbb{Z}\}$ of class $\mathcal{A}^+$ is

$$\mathrm{E}\{Y_j(1-\tau)\psi_k[Y_k(1)]\} = 0, \qquad j,\ k = 1, \ldots, K,\ \tau \geq 0,\ j \neq k \text{ or } \tau > 0 \quad (35)$$

where, in both cases, ψ_k denotes the score function of the density of $Y_k(t)$.

A similar result, concerning the contrast (26), can be obtained by a combination of the proofs of Propositions 5 and 3.

Proposition 6: Assume that the sources are filtered $(m-1)$th-order Markov processes. A necessary condition for the contrast (26) to be minimized at the sequence $\{\boldsymbol{B}(l), l \in \mathbb{Z}\}$ of class $\mathcal{A}$ is that

$$\mathrm{E}\{[Y_j(1-\tau)\cdots Y_j(m-\tau)]\psi^*_{k,m}[Y_k(1), \ldots, Y_k(m)]\} = 0, \qquad j,\ k = 1, \ldots, K,\ \tau \in \mathbb{Z},\ j \neq k \text{ or } \tau \neq 0 \quad (36)$$

where $\psi^*_{k,m}$ is as in Proposition 3.

For the instantaneous mixtures case, the estimating equations associated with (27) and (28) are somewhat more complex.

Proposition 7: Assume that the source processes are linear. If $\boldsymbol{B}$ minimizes (27) and the infimum of

$$h[(b_k \star Y_k)(1)] - \int_{-\pi}^{\pi} \log\left|\sum_{l=-\infty}^{\infty} b_k(l)e^{il\lambda}\right| d\lambda/(2\pi)$$

is attained at some sequence $\{b^*_k(l),\ l \in \mathbb{Z}\}$ of class $\mathcal{A}$, then

$$\mathrm{E}\{(b^*_k \star Y_j)(1)\psi_k[(b^*_k \star Y_k)(1)]\} = 0, \quad 1 \leq j \neq k \leq K \quad (37)$$

and if the sources are also causal with minimum phase and $\boldsymbol{B}$ minimizes (25) and the infimum of $h[\sum_{l=0}^{\infty} b_k(l)Y_k(1-l)]$, among all sequences $\{b_k(l), l \in \mathbb{Z}\}$ of class $\mathcal{A}^+$ with $b_k(0) = 1$, is attained at some sequence $\{b^*_k(l), l \in \mathbb{Z}\}$, then (37) is again satisfied. In both cases, ψ_k denotes the score function of the density of $(b^*_k \star Y_k)(1)$.

C. Discussion

It can be seen from the preceding results that, in the instantaneous mixtures case, the estimating equations (29)–(32) are of the form

$$\mathrm{E}\{[Y_j(1)\cdots Y_j(m)]\varphi_{k,m}[Y_k(1), \ldots, Y_k(m)]\} = 0, \qquad 1 \leq j \neq k \leq K \quad (38)$$

where $\varphi_{k,m}$ is a function from $\mathbb{R}^m$ to $\mathbb{R}^m$. The estimating equations (33) associated with the Gaussian contrast C_g can be viewed as a limiting form of the above, as shown before. This is also true for the estimating equations (37) associated with the contrasts (27) and (28), as they can be put into the form (38) if one truncates the sequence $\{b^*_k(l),\ l \in \mathbb{Z}\}$ to a finite sequence (which one must in practice). In this case, $\varphi_{k,m}$ takes the form

$$\varphi_{k,m}(y(1), \ldots, y(m)) = \begin{bmatrix} \beta_{k,1} \\ \vdots \\ \beta_{k,m} \end{bmatrix} \varphi_k\left[\sum_{l=1}^{m} b_{k,l}y(l)\right] \quad (39)$$

for some real function φ_k of a real variable and some real numbers $\beta_{k,1}, \ldots, \beta_{k,m}$. In all cases, the functions $\varphi_{k,m}$, referred to as separating functions, are related to the densities of the sources in a specific way (they are linear in the Gaussian case).

Turning to the convolutive mixture case, we see that the estimating equations (34)–(36) are of the form

$$\mathrm{E}\{[Y_j(1-\tau)\cdots Y_j(m-\tau)]\varphi_{k,m}[Y_k(1), \ldots, Y_k(m)]\} = 0, \qquad j,\ k = 1, \ldots, K,\ \tau \in \mathbb{Z},\ j \neq k \text{ or } \tau \neq 0 \quad (40)$$

where $\varphi_{k,m}$ is a function from $\mathbb{R}^m$ to $\mathbb{R}^m$, with $m = 1$ in the case of linear sources and τ constrained to be nonnegative in the case where the sources are further causal with minimum phase and the reconstruction sequence of matrices $\{\boldsymbol{B}(l),\ l \in \mathbb{Z}\}$ is restricted to the class $\mathcal{A}^+$. Note that the system (40) contains an infinite number of equations with an infinite number of unknowns! In practice, one may restrict $\boldsymbol{B}(l)$ to be zero for l outside some given range $[L_1, L_2]$ and restrict τ to the same range, so as to have just K equations less than the number of unknowns (which accounts for the indeterminacy of scale). Note, however, that taking $L_1 = 0$ is not enough to ensure that the sequence $\{\boldsymbol{B}(l), l \in \mathbb{Z}\}$ is of class $\mathcal{A}^+$. This constraint is actually not easy to enforce.

The use of a system of estimating equations of the form (38) or (40) is much more flexible than that of contrasts, since such a system *needs not arise from the differentiation of a contrast*. In the context of blind source separation, it is simply a system of equations, which is satisfied when the reconstructed sources are independent [5]. (Note that the system generally includes the expectation operator, which should be replaced by appropriate sample average before being solved to obtain the estimates of the parameters.) It can be easily seen that any system of the form (38) or (40) is a system of estimating equations, as soon as the sources (or the $\varphi_{k,m}[Y_k(1), \ldots, Y_k(m)]$) have

zero mean. Note that taking $m = 1$ in (38) yields the set of estimating functions introduced in [9], which can be traced back to the method in [7], while taking $\varphi_{k,m}$ of the form (39) with φ_k being the identity function yields the method for separating correlated sources in [9]. Also, many *ad hoc* methods for blind source separation consist in equating to zero the cross cumulants of higher order, possibly with lag, between the reconstructed sources. This amounts roughly to solving a system of the form (38) or (40).

But there is a price to pay for the above flexibility. First, the system of estimating equations constitutes only a necessary condition, it can (and often does) lead to spurious reconstructed sources, as such equations often have multiple solutions. We believe that by deriving them from a contrast one has a better chance of avoiding this problem. The fact that they come from a contrast makes it possible to monitor the calculation algorithm so as to ensure that the contrast is decreased at each step of the algorithm (and thus the reconstructed sources are closer to independence after each step in some sense). Second, the choice of the separating functions $\varphi_{k,m}$ can have great impact on the performance of the method: a bad choice could severely degrade the performance. Our results provide a set of good candidates for the separating functions, as they are derived from the mutual information contrast, which is related to the maximum-likelihood principle (see [2]). These functions need not be exactly the ones given in our propositions though. They can be simply some rough estimates of them. Note that the general form (38) or (40) requires the specification of Km real functions of m real variables and thus allows many degrees of freedom. If one believes that the sources are linear processes, or may be well approximated by such processes, one may settle for separating functions of the form (39) or (40) with $m = 1$, which requires the specification of only K real functions, but K linear filters need also to be estimated or specified.

APPENDIX
PROOFS OF RESULTS

A. Proof of Lemma 1

By (7)

$$h[Y(t)|Y(t-1), \ldots, Y(1)] \geq h[Y(m)|Y(m-1), \ldots, Y(1)], \quad 1 \leq t < m. \quad (41)$$

Therefore, from (5) one gets the inequality of the lemma. On the other hand, by (5) again

$$h[Y(1), \ldots, Y(m)]/m - h[Y(1), \ldots, Y(m-1)]/(m-1) = \frac{1}{m} h[Y(m)|Y(m-1), \ldots, Y(1)] - \frac{1}{m(m-1)} \sum_{t=1}^{m-1} h[Y(t)|Y(t-1), \ldots, Y(1)].$$

But from (41) the above right-hand side is nonpositive. Thus, $h[Y(1), \ldots, Y(m)]/m$ is nonincreasing in m, as well as $h[Y(m)|Y(m-1), \ldots, Y(1)]$, as implied by (7). This implies their convergence (a result already proved in [4]). □

B. Proof of Proposition 1

The proof relies on the following result.

Lemma 6: Let $\{\boldsymbol{X}(t), t \in \mathbb{Z}\}$ be a vector random stationary process admitting αth absolute moment ($\alpha \geq 1$) and $\boldsymbol{Y}(t) = (\boldsymbol{B} \star \boldsymbol{X})(t)$ where $\{\boldsymbol{B}(l), l \in \mathbb{Z}\}$ is a sequence of matrices with only a finite number of nonzero terms. Then

$$h[\boldsymbol{Y}(\cdot)] \geq h[\boldsymbol{X}(\cdot)] + \int_{-\pi}^{\pi} \log \left| \det \sum_{j=-\infty}^{\infty} \boldsymbol{B}(l) e^{il\lambda} \right| \frac{d\lambda}{2\pi}.$$

Proof: Consider the random vectors

$$\boldsymbol{Y}^{(m)}(t) = \sum_{l=-\infty}^{\infty} \boldsymbol{B}(l) \boldsymbol{X}[(t-l) \pmod m] = \sum_{\tau=0}^{m-1} \left[\sum_{l=-\infty}^{\infty} \boldsymbol{B}(t-\tau+lm) \right] \boldsymbol{X}(\tau).$$

This defines a linear transformation from $\boldsymbol{X}(0), \ldots, \boldsymbol{X}(m-1)$ to $\boldsymbol{Y}^{(m)}(0), \ldots, \boldsymbol{Y}^{(m)}(m-1)$ with the transformation matrix $\boldsymbol{B}^{(m)}$ being block circular Toeplitz with $\sum_{l=-\infty}^{\infty} \boldsymbol{B}(t-\tau+lm)$ at the (τ, t) place. Thus, by (12)

$$h[\boldsymbol{Y}^{(m)}(0), \ldots, \boldsymbol{Y}^{(m)}(m-1)] = h[\boldsymbol{X}(0), \ldots, \boldsymbol{X}(m-1)] + \log |\det \boldsymbol{B}^{(m)}|.$$

To compute the determinant of $\boldsymbol{B}^{(m)}$, note that if $\boldsymbol{U}$ is the block matrix with $e^{-i2\pi t\tau/m} \boldsymbol{I}/\sqrt{m}$ at the (τ, t) place, then $\boldsymbol{U}$ is unitary and $\boldsymbol{U}^{-1}\boldsymbol{B}^{(m)}\boldsymbol{U}$ is block diagonal with diagonal blocks $\sum_{l=-\infty}^{\infty} e^{i2\pi lt/m} \boldsymbol{B}(l)$, $t = 0, \ldots, m-1$. Therefore,

$$\log |\det \boldsymbol{B}^{(m)}| = \sum_{t=0}^{m-1} \log \left| \det \sum_{l=-\infty}^{\infty} e^{i2\pi lt/T} \boldsymbol{B}(l) \right|.$$

On the other hand, by assumption, $\boldsymbol{B}(l) = 0$ as soon as $|l|$ is greater than some integer, say q. Then it is easily seen that for $m > 2q$, $\boldsymbol{Y}^{(m)}(t) = \boldsymbol{Y}(t)$ if $t \in \{q, \ldots, m-1-q\}$. Thus, since the mutual information is nonnegative

$$h[\boldsymbol{Y}^{(m)}(0), \ldots, \boldsymbol{Y}^{(m)}(m-1)] \leq h[\boldsymbol{Y}(q), \ldots \boldsymbol{Y}(m-1-q)] + \sum_{0 \leq t < q,\, m-q \leq t < m} h[\boldsymbol{Y}^{(m)}(t)].$$

By assumption, the random vectors $\boldsymbol{Y}^{(m)}(t)$ admit αth absolute moment bounded by a constant not depending on m and t. We shall show in what follows that for a random vector Z with bounded αth absolute moment, $h(Z)$ is bounded above, regardless of the density of Z. From this and the preceding results

$$h[\boldsymbol{Y}(q), \ldots, \boldsymbol{Y}(m-1-q)] \geq h[\boldsymbol{X}(0), \ldots, \boldsymbol{X}(m-1)] + \sum_{t=0}^{m-1} \log \left| \det \sum_{l=-\infty}^{\infty} e^{i2\pi lt/m} \boldsymbol{B}(l) \right| - C$$

C being a constant not depending on m. Dividing both sides of the preceding inequality by m, then letting $m \to \infty$, one gets the result of the lemma.

To complete the proof, we need to show the assertion mentioned earlier. Let p_Z be the density of Z and put

$q(z) = C_\alpha e^{-\|z\|^\alpha}$ where $\|\cdot\|$ denotes a vector norm and C_α is the normalizing constant so that q is a density. Then

$$h(Z) = \mathrm{E}\log[q(Z)/p_Z(Z)] - \mathrm{E}\log q(Z) \le \mathrm{E}[q(Z)/p_Z(Z) - 1] + \mathrm{E}\|Z\|^\alpha - \log C_\alpha.$$

But the first term on the last right-hand side vanishes because q is a density, yielding the announced result. □

Proof of the Proposition: Let m, ϵ, and δ be as in Definition 2 and $\{\boldsymbol{B}^\dagger(l), l \in \mathbb{Z}\}$ be the inverse sequence (with respect to the convolution) of $\{\boldsymbol{B}(l), l \in \mathbb{Z}\}$. Write $\boldsymbol{B}(l) = \boldsymbol{B}_n(l) + \overline{\boldsymbol{B}}_n(l)$ where $\overline{\boldsymbol{B}}_n(l) = \boldsymbol{B}(l)$ if $|l| \le n$, $= \boldsymbol{0}$ otherwise. Then

$$\boldsymbol{Y}_n(t) = (\boldsymbol{B}_n \star \boldsymbol{X})(t) = \boldsymbol{Y}(t) - (\overline{\boldsymbol{B}}_n \star \boldsymbol{B}^\dagger \star \boldsymbol{Y})(t).$$

Therefore, since

$$\sum_{l=-\infty}^{\infty} \|(\overline{\boldsymbol{B}}_n \star \boldsymbol{B}^\dagger)(l)\| \le \left[\sum_{|l|>n} \|\overline{\boldsymbol{B}}(l)\|\right]\left[\sum_{l=-\infty}^{\infty} \|\boldsymbol{B}^\dagger(l)\|\right]$$

one can choose n sufficiently large such that it is bounded by δ, and the continuity condition of Proposition 1 entails that $h[\boldsymbol{Y}_n(1), \ldots, \boldsymbol{Y}_n(m)] \le h[\boldsymbol{Y}(1), \ldots, \boldsymbol{Y}(m)] + \epsilon$. Therefore, by Lemma 1, $h[\boldsymbol{Y}(1), \ldots, \boldsymbol{Y}(m)] + \epsilon \ge mh[\boldsymbol{Y}_n(\cdot)]$. Applying now Lemma 6 to the process $\{\boldsymbol{Y}_n(t), t \in \mathbb{Z}\}$, one gets

$$h[\boldsymbol{Y}(1), \ldots, \boldsymbol{Y}(m)]/m + \epsilon/m \ge h[\boldsymbol{X}(\cdot)] + \int_{-\pi}^{\pi} \log\left|\det \sum_{j=-\infty}^{\infty} \boldsymbol{B}_n(l) e^{il\lambda}\right| \frac{d\lambda}{2\pi}.$$

Letting $n \to \infty$ and then $m \to \infty$, one gets the same inequality as in Lemma 6. But since $\boldsymbol{X}(t) = (\boldsymbol{B}^\dagger \star \boldsymbol{Y})(t)$, one may apply the result just proved and obtains the reverse inequality

$$h[\boldsymbol{X}(\cdot)] \ge h[\boldsymbol{Y}(\cdot)] + \int_{-\pi}^{\pi} \log\left|\det \sum_{j=-\infty}^{\infty} \boldsymbol{B}^\dagger(l) e^{il\lambda}\right| \frac{d\lambda}{2\pi} = h[\boldsymbol{Y}(\cdot)] - \int_{-\pi}^{\pi} \log\left|\det \sum_{j=-\infty}^{\infty} \boldsymbol{B}(l) e^{il\lambda}\right| \frac{d\lambda}{2\pi}.$$

It follows that the last inequality is an equality. □

C. Proof of Lemma 2

For convenience, put

$$\boldsymbol{Y} = [\boldsymbol{Y}^{\mathrm{T}}(1) \cdots \boldsymbol{Y}^{\mathrm{T}}(m)]^{\mathrm{T}}$$
$$\boldsymbol{Y}' = [(\boldsymbol{C} * \boldsymbol{Y})^{\mathrm{T}}(1) \cdots (\boldsymbol{C} * \boldsymbol{Y})^{\mathrm{T}}(m)]^{\mathrm{T}}$$

and denote by $p_{\boldsymbol{Y}}$, $p_{\boldsymbol{Y}'}$ their densities. Then

$$h(\boldsymbol{Y}') - h(\boldsymbol{Y}) = \mathrm{E}[\log p_{\boldsymbol{Y}}(\boldsymbol{Y}) - \log p_{\boldsymbol{Y}}(\boldsymbol{Y}')] - \mathrm{E}\log[p_{\boldsymbol{Y}'}(\boldsymbol{Y}')/p_{\boldsymbol{Y}}(\boldsymbol{Y}')].$$

The last term is nonnegative since it is a Kullback–Leibler divergence. As for the first term, by the mean value theorem and our assumption and the fact that

$$(a+b)^\lambda \le 2^{\max(\lambda-1,\,0)}(a^\lambda + b^\lambda)$$

for positive a, b, λ, one has

$$|\log p_{\boldsymbol{Y}}(\boldsymbol{Y}) - \log p_{\boldsymbol{Y}}(\boldsymbol{Y}')| \le C[1 + 2^{\max(\alpha-2,\,0)}(\|\boldsymbol{Y}\|^{\alpha-1} + \|\boldsymbol{Y}' - \boldsymbol{Y}\|^{\alpha-1})]\|\boldsymbol{Y}' - \boldsymbol{Y}\|.$$

But by the Hölder inequality, the last expression has expectation bounded by a constant times $\|\boldsymbol{C}(0) - \boldsymbol{I}\| + \sum_{l\ne 0}\|\boldsymbol{C}(l)\|$, which yields the result. □

D. Proof of Lemma 3

Let $p_{Y_k(m)|Y_k(m-1),\ldots,Y_k(1)}$ and $p_{\boldsymbol{Y}(m)|\boldsymbol{Y}(m-1),\ldots,\boldsymbol{Y}(1)}$ denote the conditional densities of $Y_k(m)$ given $Y_k(m-1), \ldots, Y_k(m-1)$ and of $\boldsymbol{Y}(m)$ given $\boldsymbol{Y}(m-1), \ldots, \boldsymbol{Y}(1)$. Then the expected divergence mentioned in Lemma 3 is

$$-\mathrm{E}\log \frac{\prod_{k=1}^{K} p_{Y_k(m)|Y_k(m-1),\ldots,Y_k(1)}[Y_k(1), \ldots, Y_k(m)]}{p_{\boldsymbol{Y}(m)|\boldsymbol{Y}(m-1),\ldots,\boldsymbol{Y}(1)}[\boldsymbol{Y}_k(1), \ldots, \boldsymbol{Y}_k(m)]}.$$

But this expression can be easily seen to be

$$\sum_{k=1}^{K} h[Y_k(m)|Y_k(m-1), \ldots, Y_k(1)] - h[\boldsymbol{Y}(m)|\boldsymbol{Y}(m-1), \ldots, \boldsymbol{Y}(1)]$$

and the last term (without the minus sign), by (6) and (12), equals $h[\boldsymbol{X}(m)|\boldsymbol{X}(m-1), \ldots, \boldsymbol{Y}(1)] - \log\det\boldsymbol{B}$, yielding the result. □

E. Proof of Proposition 2

Let $\{\boldsymbol{D}(l), l \in \mathbb{Z}\}$ be any sequence of diagonal matrices of class $\mathcal{A}$ and put $\tilde{\boldsymbol{B}}(l) = (\boldsymbol{D} \star \boldsymbol{B})(l)$, $\tilde{Y}_k(t) = (d_k \star Y_k)(t)$ where $d_k(l)$ are the diagonal elements of $\boldsymbol{D}(l)$. Then one can write

$$C_1[\boldsymbol{B}(\cdot)] = C_\infty\left[\tilde{\boldsymbol{B}}(\cdot)\right] + \sum_{k=1}^{K}\left\{h[Y_k(1)] - h\left[\tilde{Y}_k(\cdot)\right] + \int_{-\pi}^{\pi} \log\left|\det \sum_{l=-\infty}^{\infty} d_k(l) e^{il\lambda}\right| \frac{d\lambda}{2\pi}\right\}.$$

Each term in the sum in the above right-hand side is nonnegative since, by Corollary 2

$$h\left[\tilde{Y}_k(\cdot)\right] \le h[Y_k(1)] + \int_{-\pi}^{\pi} \log\left|\sum_{l=-\infty}^{\infty} d_k(l) e^{il\lambda}\right| \frac{d\lambda}{2\pi}. \quad (42)$$

Further, $C_\infty(\tilde{\boldsymbol{B}})$ is no other than $I[\tilde{Y}_1(\cdot), \ldots, \tilde{Y}_k(\cdot)] + h[\boldsymbol{X}(\cdot)]$, hence, $C_1[\boldsymbol{B}(\cdot)]$ is bounded below by (the constant) $h[\boldsymbol{X}(\cdot)]$. This bound will be attained if the processes $\{Y_k(t), t \in \mathbb{Z}\}$ are temporally independent and independent among themselves and it is possible to choose $\{\boldsymbol{B}(l), l \in \mathbb{Z}\}$ so that these processes are so, since the sources are independent linear processes and the sequences $\{a_k(l), l \in \mathbb{Z}\}$ in their representation (23) are of class $\mathcal{A}$.

It remains to show that $C_1[\boldsymbol{B}(\cdot)]$ can attain its minimum only if the processes $\{Y_k(t), t \in \mathbb{Z}\}$ are temporally independent and independent among themselves. We observe that, by Corollary 2, (42) can be an equality only if the process $\{Y_k(t), t \in \mathbb{Z}\}$ is temporally independent. Thus, $C_1[\boldsymbol{B}(\cdot)]$ can attain its minimum only if this happens for all k and $I[\tilde{Y}_1(\cdot), \ldots, \tilde{Y}_k(\cdot)] = 0$. But

by Proposition 1, $I[\tilde{Y}_1(\cdot), \ldots, \tilde{Y}_k(\cdot)] = I[Y_1(\cdot), \ldots, Y_k(\cdot)]$ and because the processes $\{Y_k(t), t \in \mathbb{Z}\}$ are temporally independent

$$I[Y_1(\cdot), \ldots, Y_k(\cdot)] = \frac{1}{m} I\left\{[Y_1(1)\cdots Y_1(m)]^{\mathrm{T}}, \ldots, [Y_K(1)\cdots Y_K(m)]^{\mathrm{T}}\right\}$$

for all m. Thus, the last right-hand side vanishes for all m and hence the processes $\{Y_k(t), t \in \mathbb{Z}\}$ are independent. □

F. Proof of Lemma 4

Put $\boldsymbol{Y}' = \boldsymbol{Y} + \mathcal{E}\boldsymbol{Z}$, then by the same calculation as in the proof of Lemma 2 and using C1)

$$h(\boldsymbol{Y}') - h(\boldsymbol{Y}) = \mathrm{E}[\log p_{\boldsymbol{Y}}(\boldsymbol{Y}) - \log p_{\boldsymbol{Y}}(\boldsymbol{Y}')] + o(\mathcal{E}).$$

Therefore, one gets the result of the proposition if one has proved that

$$\mathrm{E}\left\{[\log p_{\boldsymbol{Y}}(\boldsymbol{Y}) - \log p_{\boldsymbol{Y}}(\boldsymbol{Y}') - \psi_{\boldsymbol{Y}}^{\mathrm{T}}(\boldsymbol{Y})\mathcal{E}\boldsymbol{Z}]/\|\mathcal{E}\|\right\} \to 0, \quad \text{as } \|\mathcal{E}\| \to 0.$$

Since the random variable inside the curly bracket converges to 0 almost surely as $\|\mathcal{E}\| \to 0$, by the Lebesgue dominated convergence theorem, one needs only to show that it is bounded for all $\mathcal{E}$ small enough by a fixed integrable random variable. But, repeating again the argument in the proof of Lemma 2, this random variable is bounded by

$$C[2 + 2^{\max(\alpha-2,0)}(\|\boldsymbol{Y}\|^{\alpha-1} + c\|\boldsymbol{Z}\|^{\alpha-1}) + \|\boldsymbol{Y}\|^{\alpha-1}]\|\boldsymbol{Z}\|.$$

for all $\mathcal{E}$ such that $\|\mathcal{E}\| \leq c$. The last random variable is integrable by the Hölder inequality, yielding the result. □

G. Proof of Proposition 3

For C_m to be minimized at $\boldsymbol{B}$ one must have $C_m(\boldsymbol{B}+\mathcal{E}\boldsymbol{B}) \geq C_m(\boldsymbol{B})$ for all matrices $\mathcal{E}$. Take $\mathcal{E}$ having a single nonzero element, say $\mathcal{E}_{kj}$, and put $\boldsymbol{Y}_k = [Y_k(1)\cdots Y_k(m)]^{\mathrm{T}}$. One gets from (18)

$$[h(\boldsymbol{Y}_k + \mathcal{E}_{kj}\boldsymbol{Y}_j) - h(\boldsymbol{Y}_k)]/m - \log|\det(\boldsymbol{I} + \mathcal{E})| \geq 0.$$

But by Proposition 4, the first term in the preceding left-hand side equals $\mathcal{E}_{kj}\mathrm{E}[\boldsymbol{Y}_j\psi_{k,m}(\boldsymbol{Y}_k)]/m + o(\mathcal{E}_{kj})$ and if $k \neq j$, $\det(\boldsymbol{I} + \mathcal{E})| = 1$. Therefore, for $\mathcal{E}_{kj}$ small enough

$$\mathcal{E}_{kj}\mathrm{E}[\boldsymbol{Y}_j\psi_{k,m}(\boldsymbol{Y}_k)] \geq 0$$

and since $\mathcal{E}_{kj}$ is arbitrary, one must have $\mathrm{E}[\boldsymbol{Y}_j\psi_{k,m}(\boldsymbol{Y}_k)] = 0$, which yields (29).

Since $C_m^* = mC_m - (m-1)C_{m-1}$, a completely similar argument yields that a necessary condition for it to be minimized at $\boldsymbol{B}$ is that (30) holds with

$$\psi_{k,m}^*[y(1), \ldots, y(m)] = \psi_{k,m}[y(1), \ldots, y(m)] - \begin{bmatrix} \psi_{k,m-1}[y(1), \ldots, y(m-1)] \\ 0 \end{bmatrix}.$$

But one can easily see that $\psi_{k,m}^*$ as defined here is the same as the one in the proposition. □

H. Proof of Proposition 4

Observe that

$$\log|\det(\boldsymbol{M} + \boldsymbol{\delta})| = \mathrm{tr}(\boldsymbol{M}^{-1}\boldsymbol{\delta}) + o(\|\boldsymbol{\delta}\|)$$

as $\boldsymbol{\delta} \to \boldsymbol{0}$, tr denoting the trace. Hence, when $\boldsymbol{B}$ is changed to $\boldsymbol{B} + \mathcal{E}\boldsymbol{B}$ where $\mathcal{E}$ is a matrix with only a nonzero term $\mathcal{E}_{kj}$ with $j \neq k$, one gets from (21) that the change of $C_{g,m}$ is

$$\mathcal{E}_{kj}\mathrm{tr}[\mathrm{cov}(\boldsymbol{Y}_k)^{-1}\mathrm{cov}(\boldsymbol{Y}_j, \boldsymbol{Y}_k)]/m + o(\mathcal{E}_{kj})$$

where we have put $\boldsymbol{Y}_k = [Y_k(1)\cdots Y_k(m)]^{\mathrm{T}}$ while $\mathrm{cov}(\cdot)$ and $\mathrm{cov}(\cdot, \cdot)$ refer to the covariance and cross covariance matrices, respectively. For $\boldsymbol{B}$ to maximize $C_{g,m}$, it is necessary that the last expression be nonpositive for all $\mathcal{E}_{kj}$. Therefore,

$$\mathrm{tr}[\mathrm{cov}(\boldsymbol{Y}_k)^{-1}\mathrm{cov}(\boldsymbol{Y}_j, \boldsymbol{Y}_k)] = 0$$

which is no other than the condition (31).

On the other hand, by (22), the change of $C_{g,m}^*$ corresponding to the same above change of $\boldsymbol{B}$, is, putting $e_k(m) = Y_k(m) - Y_k(m|1 : m-1) = \sum_{l=0}^{m-1} a_{k,m-1}(l)Y_k(m-l)$

$$\mathrm{cov}\left\{e_k(m), \sum_{l=1}^{m-1}[a_{k,m-1}(l)\mathcal{E}_{kj}Y_j(m-l) + \delta_{k,m-1}(l)Y_k(m-l)]\right\} \Big/ \mathrm{var}\{e_k(m)\} + o(\mathcal{E}_{kj})$$

where $\delta_{k,m-1}(l)$ represents the change of $a_{k,m-1}(l)$ when $Y_k(t)$ is changed to $Y_k(t) + \mathcal{E}_{kj}Y_j(t)$. But the terms involving the $\delta_{k,m-1}(l)$ disappear since $\mathrm{cov}\{e_k(m), Y_k(m-l)\} = 0$ by the definition of $Y_k(m|1 : m-1)$. Therefore, by the same argument as before, one obtains (32).

Finally, with the same change of $\boldsymbol{B}$ as before, the process $\{Y_k(t), t \in \mathbb{Z}\}$ is changed to $\{Y_k(t) + \mathcal{E}_{kj}Y_j(t), t \in \mathbb{Z}\}$ while the processes $\{Y_i(t), t \in \mathbb{Z}\}$, $i \neq k$, are unchanged. It then follows from (20) that the corresponding change of C_g^* is

$$\mathcal{E}_{kj}\int_{-\pi}^{\pi}[f_{Y_kY_j}(\lambda)/f_{Y_kY_k}(\lambda)]d\lambda/(2\pi) + o(\mathcal{E}_{kj}).$$

This yields (33). □

I. Proof of Proposition 5

Suppose that the contrast (34) is minimized at $\{\boldsymbol{B}(l), l \in \mathbb{Z}\}$, then adding to this sequence the sequence $\{(\mathcal{E} \star \boldsymbol{B})(l), l \in \mathbb{Z}\}$, where $\{\mathcal{E}(l), l \in \mathbb{Z}\}$ is any sequence of matrices, must not decrease this contrast. Choose this sequence to have only one nonzero term $\mathcal{E}(\tau)$ which has only one nonzero element $\mathcal{E}_{kj}(\tau)$, then the change of this contrast is

$$h[Y_k(1) + \mathcal{E}_{kj}(\tau)Y_j(1-\tau)] - h[Y_k(t)] - \int_{-\pi}^{\pi}\log|\det[\boldsymbol{I} + \mathcal{E}(\tau)e^{i\tau\lambda}]|\frac{d\lambda}{2\pi}.$$

If either $\tau \neq 0$ or $j \neq k$, the last term of the above expression is $o[\mathcal{E}_{kj}(\tau)]$ as $\mathcal{E}_{kj}(\tau) \to 0$. As for the contribution of the other terms, using Proposition 4, it can be seen to be $\mathcal{E}_{kj}(m)\mathrm{E}\{Y_j(1-\tau)\psi_k[Y_k(1)]\} + o[\mathcal{E}_{kj}(\tau)]$. This yields the first result of the proposition. The proof for the other result is similar. □

J. Proof of Proposition 7

For the contrast (27), when $\boldsymbol{B}$ is changed to $\boldsymbol{B}+\mathcal{E}\boldsymbol{B}$ with $\mathcal{E}$ having a single nonzero element $\mathcal{E}_{kj}$, the corresponding change of this contrast is the infimum of

$$h[(b_k \star Y_k)(1)+\mathcal{E}_{kj}(b_k \star Y_j)(1)] - h[(b_k^* \star Y_k)(1)] + \int_{-\pi}^{\pi} \log \left| \frac{\sum_{l=-\infty}^{\infty} b_k^*(l)e^{il\lambda}}{\sum_{l=-\infty}^{\infty} b_k(l)e^{il\lambda}} \right| \frac{d\lambda}{2\pi} - \log|\det(\boldsymbol{I}+\mathcal{E})|$$

over all sequences $\{b_k(l), l \in \mathbb{Z}\}$ of class $\mathcal{A}$. As $\boldsymbol{B}$ minimizes the contrast, the above expression must be nonnegative; hence, taking $j \neq k$ (so that $\det(\boldsymbol{I}+\mathcal{E})=1$) and $b_k(l)=b_k^*(l)$

$$h[(b_k^* \star Y_k)(t)+\mathcal{E}_{kj}(b_k^* \star Y_j)(t)] - h[(b_k^* \star Y_k)(t)] \geq 0.$$

The above right-hand side, by Proposition 4, is

$$\mathrm{E}\{(b_k^* \star Y_j)(1)\psi_k[(b_k^* \star Y_k)(1)]\} + o(\mathcal{E}_{kj})$$

yielding the first result of the proposition. The proof for the other result is similar. □

REFERENCES

[1] A. Belouchrani, K. Abed-Meraim, J. F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Trans. Signal Processing*, vol. 45, pp. 434–444, Feb. 1997.

[2] J.-F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 86, pp. 2009–2026, Aug. 1998.

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

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

[5] V. P. Godambe, "Conditional likelihood and unconditional optimum estimating equations," *Biometrika*, vol. 63, pp. 277–284, 1963.

[6] E. J. Hannan, *Multiple Time Series*. New York: Wiley, 1970.

[7] C. Jutten and J. Hérault, "Blind separation of sources, Part I: An adaptative algorithm based on neuromimetic structure," *Signal Processing*, vol. 24, pp. 1–10, 1991.

[8] D. T. Pham, "Blind separation of instantaneous mixture of sources via an independent component analysis," *IEEE Trans. Signal Processing*, vol. 44, pp. 2768–2779, Nov. 1996.

[9] D. T. Pham and P. Garat, "Blind separation of mixtures of independent sources through a quasi maximum likelihood approach," *IEEE Trans. Signal Processing*, vol. 45, pp. 1712–1725, July 1997.

[10] D. T. Pham, "Mutual information approach to blind separation of sources," in *Proc. ICA'99 Workshop*, Aussois, France, Jan. 1999, pp. 215–220.

[11] ——, "Blind separation of instantaneous mixture of sources via the Gaussian mutual information criterion," *Signal Processing*, vol. 81, pp. 855–870, 2001.

[12] ——, Contrast functions for ICA and sources separation. Tech. Rep.. [Online]. Available: http://www-lmc.imag.fr/lmc-sms/Dinh-Tuan.Pham/BSS/contrast.ps.gz

[13] ——, "Contrast functions for blind separation and deconvolution of sources," in *Proc. ICA 2001 Conf.*, San Diego, CA, Dec. 2001.

[14] A. Zygmund, *Trigonometric Series*. Cambridge, U.K.: Cambridge Univ. Press, 1968, vol. I.

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

He was a Postdoctoral Fellow in the Department of Statistics, University of California, Berkeley during 1977–1978 and a Visiting Professor in the Department of Mathematics, University of Indiana at Bloomingtion during 1979–1980. He is currently Director of Researche at the French Centre National de Recherche Scientifique, Grenoble. His researches includes time-series analysis, signal modeling, blind source separation, array processing, and biomedical signal processing.