Vol.36, no 3, Special issue on High-Order Statistics, April 1994

ELSEVIER

Signal Processing 36 (1994) 287–314

SIGNAL PROCESSING

Independent component analysis, A new concept?†

Pierre Comon

THOMSON-SINTRA, Parc Sophia Antipolis, BP 138, F-06561 Valbonne Cedex, France

Received 24 August 1992

Abstract

The independent component analysis (ICA) of a random vector consists of searching for a linear transformation that minimizes the statistical dependence between its components. In order to define suitable search criteria, the expansion of mutual information is utilized as a function of cumulants of increasing orders. An efficient algorithm is proposed, which allows the computation of the ICA of a data matrix within a polynomial time. The concept of ICA may actually be seen as an extension of the principal component analysis (PCA), which can only impose independence up to the second order and, consequently, defines directions that are orthogonal. Potential applications of ICA include data analysis and compression, Bayesian detection, localization of sources, and blind identification and deconvolution.

Zusammenfassung

Die Analyse unabhängiger Komponenten (ICA) eines Vektors beruht auf der Suche nach einer linearen Transformation, die die statistische Abhängigkeit zwischen den Komponenten minimiert. Zur Definition geeigneter Such-Kriterien wird die Entwicklung gemeinsamer Information als Funktion von Kumulanten steigender Ordnung genutzt. Es wird ein effizienter Algorithmus vorgeschlagen, der die Berechnung der ICA für Datenmatrizen innerhalb einer polynomischen Zeit erlaubt. Das Konzept der ICA kann eigentlich als Erweiterung der 'Principal Component Analysis' (PCA) betrachtet werden, die nur die Unabhängigkeit bis zur zweiten Ordnung erzwingen kann und deshalb Richtungen definiert, die orthogonal sind. Potentielle Anwendungen der ICA beinhalten Daten-Analyse und Kompression, Bayes-Detektion, Quellenlokalisierung und blinde Identifikation und Entfaltung.

Résumé

L'Analyse en Composantes Indépendantes (ICA) d'un vecteur aléatoire consiste en la recherche d'une transformation linéaire qui minimise la dépendance statistique entre ses composantes. Afin de définir des critères d'optimisation appropriés, on utilise un développement en série de l'information mutuelle en fonction de cumulants d'ordre croissant. On propose ensuite un algorithme pratique permettant le calcul de l'ICA d'une matrice de données en un temps polynomial. Le concept d'ICA peut être vu en réalité comme une extension de l'Analyse en Composantes Principales (PCA) qui, elle, ne peut imposer l'indépendance qu'au second ordre et définit par conséquent des directions orthogonales. Les applications potentielles de l'ICA incluent l'analyse et la compression de données, la détection bayesienne, la localisation de sources, et l'identification et la déconvolution aveugles.

† This work was supported in part by the DRET.

SSDI 0165-1684(93)E0093-Z

Key words: Statistical independence; Random variables; Mixture, Entropy; Information; High-order statistics; Source separation; Data analysis; Noise reduction; Blind identification; Array processing; Principal components; Independent components

1. Introduction

1.1. Problem description

This paper attempts to provide a precise definition of ICA within an applicable mathematical framework. It is envisaged that this definition will provide a baseline for further development and application of the ICA concept. Assume the following linear statistical model:

$$y = Mx + v, \tag{1.1a}$$

where x, y and v are random vectors with values in $\mathbb{R}$ or $\mathbb{C}$ and with zero mean and finite covariance, M is a rectangular matrix with at most as many columns as rows, and vector x has statistically independent components. The problem set by ICA may be summarized as follows. Given T realizations of y, it is desired to estimate both matrix M and the corresponding realizations of x. However, because of the presence of the noise v, it is in general impossible to recover exactly x. Since the noise v is assumed here to have an unknown distribution, it can only be treated as a nuisance, and the ICA cannot be devised for the noisy model above, Instead, it will be assumed that

$$y = Fz, \tag{1.1b}$$

where z is a random vector whose components are maximizing a 'contrast' function. This terminology is consistent with [31], where Gassiat introduced contrast functions in the scalar case for purposes of blind deconvolution. As defined subsequently, the contrast of a vector z is maximum when its components are statistically independent. The qualifiers 'blind' or 'myopic' are often used when only the outputs of the system considered are observed; in this framework, we are thus dealing with the problem of blind identification of a linear static system.

1.2. Organization of the paper

In this section, related works, possible applications and preliminary observations regarding the problem statement are surveyed. In Section 2, general results on statistical independence are stated, and are then utilized in Section 3 to derive optimization criteria. The properties of these criteria are also investigated in Section 3. Section 4 is dedicated to the design of a practical algorithm, delivering a solution within a number of steps that is a polynomial function of the dimension of vector y. Simulation results are then presented in Section 5. For the sake of clarity, most proofs as well as a discussion about complex data have been deferred to the appendix at the end of the paper.

1.3. Related works

The calculation of ICA was discussed in several recent papers [8, 16, 30, 36, 37, 61], where the problem was given various names. For instance, the terminology 'sources separation problem' has often been coined. Investigations reveal that the problem of 'independent component analysis' was actually first proposed and so named by Herault and Jutten around 1986 because of its similarities with principal component analysis (PCA). This terminology is retained in the paper.

Herault and Jutten seem to be the first (around 1983) to have addressed the problem of ICA. Several papers by these authors propose an iterative real-time algorithm based on a neuro-mimetic architecture. The authors deserve merit in their achievement of an algorithmic solution when no theoretical explanation was available at that time. Nevertheless, their solution can show lack of convergence in a number of cases. Refer to [37] and other papers in the same issue, and to [27]. In their framework, high-order statistics were not introduced explicitly. It is less well known that Bar-Ness [2]

independently proposed another approach at the same time that presented rather similar qualities and drawbacks.

Giannakis et al. [35] addressed the issue of identifiability of ICA in 1987 in a somewhat different framework, using third-order cumulants. However, the resulting algorithm required an exhaustive search. Lacoume and Ruiz [42] also sketched a mathematical approach to the problem using high-order statistics; in [29, 30], Gaeta and Lacoume proposed to estimate the mixing matrix M by maximum likelihood approach, where an exhaustive search was also necessary to determine the absolute maximum of an objective function of many variables. Thus, from a practical point of view, this was realistic only in the two-dimensional case.

Cardoso focused on the algebraic properties of the fourth-order cumulants, and interpreted them as linear operators acting on matrices. A simple case is the action on the identity yielding a cumulant matrix whose diagonalization gives an estimate of ICA [8]. When the action is defined on a set of matrices, one obtains several cumulant matrices whose *joint diagonalization* provides more robust estimates [55]. This is equivalent to optimizing a cumulant-based criterion [55], and is then similar in spirit to the approach presented herein. Other algebraic approaches, using only fourth-order cumulants, have also been investigated [9, 10].

In [36], Inouye proposed a solution for the separation of two sources, whereas at the same time Comon [13] proposed another solution for $N \geqslant 2$. Together with Cardoso's solution, these were among the first direct (within polynomial time) solutions to the ICA problem. In [61], Inouye and his colleagues derived identifiability conditions for the ICA problem. Their Theorem 2 may be seen to have connections to earlier works [19]. On the other hand, our Theorem 11 (also presented in [15, 16]) only requires pairwise independence, which is generally weaker than the conditions required by [61].

In [24], Fety addressed the problem of identifying the dynamic model of the form $y(t) = Fz(t)$, where t is a time index. In general, the identification of these models can be completed with the help of second-order moments only, and will be called the *signal* separation problem. In some cases, identifiability conditions are not fulfilled, e.g. when processes $z_n(t)$ have spectra proportional to each other, so that the *signal separation* problem degenerates into the ICA problem, where the time coherency is ignored. Independently, the *signal* separation problem has also been addressed more recently by Tong et al. [60].

This paper is an extended version of the conference paper presented at the Chamrousse workshop in July 1991 [16]. Although most of the results were already tackled in [16], the proofs were very shortened, or not stated at all, for reasons of space. They are now detailed here, within the same framework. Furthermore, some new results are stated, and complementary simulations are presented.

Among other things, it is sought to give sound justification to the choice of the objective function to be maximized. The results obtained turn out to be consistent with what was heuristically proposed in [42]. Independently of [29, 30], the author proposed to approximate the probability density by its Edgeworth expansion [16], which has advantages over Gram Charlier's expansion. Contrary to [29, 30], the hypothesis that odd cumulants vanish is not assumed here: Gaeta emphasized in [29] the consistency of the theoretical results obtained in both approaches when third-order cumulants are null.

It may be considered, however, that a key contribution of this paper consists of providing a practical algorithm that does not require an exhaustive search, but can be executed within a polynomial time, even in the presence of non-Gaussian noise. The complexity aspect is of great interest when the dimension of y is large (at least greater than 2). On the other hand, the robustness in the presence of non-Gaussian noise has not apparently been previously investigated. Various implementations of this algorithm are discussed in [12, 14, 15]; the present improved version should, however, be preferred. A real-time version can also be implemented on a parallel architecture [14].

1.4. Applications

Let us spend a few lines on related applications. If x is a vector formed of N successive time samples

of a white process, and if M is Toeplitz triangular, then model (1.1) represents nothing else but a deconvolution problem. The first column of M contains the successive samples of the impulse response of the corresponding causal filter. In the FIR case, M is also banded. Such blind identification and/or deconvolution problems have been addressed in various ways in [5, 21, 45, 46, 50, 52, 62, 64]. Note that if M is not triangular, the filter is allowed to be non-causal. Moreover, if M is not Toeplitz, the filter is allowed to be non-stationary. Thus, blind deconvolution may be viewed as a constrained ICA problem, which is not dealt with in this paper. In fact, in the present study, no structure for matrix M will be assumed, so that blind deconvolution or other constrained ICAs cannot be efficiently carried out with the help of the algorithm described in Section 4, at least in its present form.

In antenna array processing, ICA might be utilized in at least two instances: firstly, the estimation of radiating sources from unknown arrays (necessarily without localization). It has already been used in radar experimentation for example [20]. In the same spirit, the use of ICA may also have some interesting features for jammer rejection or noise reduction. Secondly, ICA might be useful in a two-stage localization procedure, if arrays are perturbed or ill-calibrated. Additional experiments need, however, to be completed to confirm the advantages of ICA under such circumstances. On the other hand, the use of high-order statistics (HOS) in this area is of course not restricted to ICA. Regarding methods for sources localization, some references may be found in [8, 9, 26, 32, 47, 53, 59]. It can be expected that a two-stage procedure, consisting of first computing an ICA, and using the knowledge of the array manifold to perform a localization in the second stage, would be more robust against calibration uncertainties. But this needs to be experimented with.

Further, ICA can be utilized in the identification of *multichannel* ARMA processes when the input is not observed and, in particular, for estimating the first coefficient of the model [12, 15]. In general, the first matrix coefficient is in fact either assumed to be known [33, 56, 62, 63], or of known form, e.g. triangular [35]. Note that the authors of [35] were the first to worry about the estimation of the first coefficient of non-monic multichannel models. A survey paper is in preparation on this subject [58].

On the other hand, ICA can be used as a data preprocessing tool before Bayesian detection and classification. In fact, by a change of coordinates, the density of multichannel data may be approximated by a product of marginal densities, allowing a density estimation with much shorter observations [17]. Other possible applications of ICA can be found in areas where PCA is of interest. The application to the analysis of chaos is perhaps one of the most exotic [25].

1.5. Preliminary observations

Suppose x is a non-degenerate (i.e. non-deterministic) Gaussian random vector of dimension ρ with statistically independent components, and z is a vector of dimension $N \geqslant \rho$ defined as $z = Cx$, where C is an $N \times \rho$ full rank matrix. Then if z has independent components, matrix C can easily be shown (see Appendix A.1) to be of the form

$$C = \Lambda Q \Delta, \tag{1.2}$$

where both Λ and Δ are real square diagonal, and Q is $N \times \rho$ and satisfies $Q^*Q = I$. If $N = \rho$, Λ is invertible and Q is unitary; this demonstrates that if both x and z have a unit covariance matrix, then C may be any unitary matrix. Theorem 11 shows that when x has at most one Gaussian component, this indetermination reduces to a matrix of the form DP, where D is a diagonal matrix with entries of unit modulus and P is permutation. This latter indetermination cannot be reduced further without additional assumptions.

DEFINITION 1. The ICA of a random vector y of size N with finite covariance V_y is a pair $\{F, \Delta\}$ of matrices such that

(a) the covariance factorizes into

$$V_y = F\Delta^2 F^*,$$

where Δ is diagonal real positive and F is full column rank ρ;

(b) the observation can be written as $y = Fz$, where z is a $\rho \times 1$ random vector with covariance Δ^2

and whose components are 'the most independent possible', in the sense of the maximization of a given 'contrast function' that will be subsequently defined.

In this definition, the asterisk denotes transposition, and complex conjugation whenever the quantity is complex (mainly the real case will be considered in this paper). The purpose of the next section will be to define such contrast functions. Since multiplying random variables by non-zero scalar factors or changing their order does not affect their statistical independence, ICA actually defines an equivalence class of decompositions rather than a single one, so that the property below holds. The definition of contrast functions will thus *have to* take this indeterminacy into account.

PROPERTY 2. If a pair $\{F, \Lambda\}$ *is an ICA of a random vector,* y, *then so is the pair* $\{F', \Lambda'\}$ *with*

$$F' = F\bar{\Lambda}DP \quad \text{and} \quad \Lambda' = P^*\bar{\Lambda}^{-1}\Lambda P,$$

where $\bar{\Lambda}$ *is a* $\rho \times \rho$ *invertible diagonal real positive scaling matrix,* D *is a* $\rho \times \rho$ *diagonal matrix with entries of unit modulus and* P *is a* $\rho \times \rho$ *permutation. For conciseness we shall sometimes refer to the matrix* $\Lambda = \bar{\Lambda}D$ *as the diagonal invertible 'scale' matrix.*

For the sake of clarity, let us also recall the definition of PCA below.

DEFINITION 3. The PCA of a random vector y of size N with finite covariance V_y is a pair $\{F, \Lambda\}$ of matrices such that

(a) the covariance factorizes into

$$V_y = F\Delta^2F^*,$$

where Δ is diagonal real positive and F is full column rank ρ;

(b) F is an $N \times \rho$ matrix whose columns are orthogonal to each other (i.e. F^*F diagonal).

Note that two different pairs can be PCAs of the same random variable y, but are also related by Property 2. Thus, PCA has exactly the same inherent indeterminations as ICA, so that we may assume the same additional arbitrary constraints [15]. In fast, for computational purposes, it is convenient to define a unique representative of these equivalence classes.

DEFINITION 4. The uniqueness of ICA as well as of PCA requires imposing three additional constraints. We have chosen to impose the following:

(c) the columns of F have unit norm;

(d) the entries of Δ are sorted in decreasing order;

(e) the entry of largest modulus in each column of F is positive real.

Each of the constraints in Definition 4 removes one of the degrees of freedom exhibited in Property 2. More precisely, (c), (d) and (e) determine $\bar{\Lambda}$, P and D, respectively. With constraint (c), the matrix F defined in the PCA decomposition (Definition 3) is unitary. As a consequence, ICA (as well as PCA) defines a so-called $\rho \times 1$ 'source vector', z, satisfying

$$y = Fz. \tag{1.3}$$

As we shall see with Theorem 11, the uniqueness of ICA requires z to have at most one Gaussian component.

2. Statements related to statistical independence

In this section, an appropriate contrast criterion is proposed first. Then two results are stated. It is proved that ICA is uniquely defined if at most one component of x is Gaussian, and then it is shown why pairwise independence is a sufficient measure of statistical independence in our problem.

The results presented in this paper hold true either for real or complex variables. However, some derivations would become more complicated if derived in the complex case. Therefore, only *real variables* will be considered most of the time for the sake of clarity, and extensions to the complex case will be mainly addressed only in the appendix. In the remaining, plain lowercase (resp. uppercase) letters denote, in general, scalar quantities (resp. tables with at least two indices, namely matrices or tensors), whereas boldface lowercase letters denote column vectors with values in $\mathbb{R}^N$.

2.1. Mutual information

Let x be a random variable with values in $\mathbb{R}^N$ and denote by $p_x(u)$ its probability density function (pdf). Vector x has mutually independent components if and only if

$$p_x(u) = \prod_{i=1}^{N} p_{x_i}(u_i). \tag{2.1}$$

So a natural way of checking whether x has independent components is to measure a distance between both sides of (2.1):

$$\delta\left(p_x, \prod_{i=1}^{N} p_{x_i} \right). \tag{2.2}$$

In statistics, the large class of f-divergences is of key importance among the possible distance measures available [4]. In these measures the roles played by both densities are not always symmetric, so that we are not dealing with proper distances. For instance, the Kullback divergence is defined as

$$\delta(p_x, p_z) = \int p_x(u) \log \frac{p_x(u)}{p_z(u)} \, du. \tag{2.3}$$

Recall that the Kullback divergence satisfies

$$\delta(p_x, p_z) \geqslant 0, \tag{2.4}$$

with equality if and only if $p_z(u) = p_z(u)$ almost everywhere. This property is due to the convexity of the logarithm [6]. Now, if we look at the form of the Kullback divergence of (2.2), we obtain precisely the average mutual information of x:

$$I(p_x) = \int p_x(u) \log \frac{p_x(u)}{\prod p_{x_i}(u_i)} \, du, \quad u \in \mathbb{C}^N. \tag{2.5}$$

From (2.4), the mutual information vanishes if and only if the variables x_i are mutually independent, and is strictly positive otherwise. We shall see that this will be a good candidate as a contrast function.

2.2. Standardization

Now denote by $\mathbb{E}^N$ the space of random variables with values in $\mathbb{R}^N$, by $\mathbb{E}_r^N$ the Euclidian subspace of $\mathbb{E}^N$ spanned by variables with finite moments up to order r, for any $r \geqslant 2$, provided with the scalar product $\langle x, y \rangle = E[x^* y]$ and by $\tilde{\mathbb{E}}_2^N$ the subset of $\mathbb{E}_2^N$ of variables having an invertible covariance. For instance, we have $\mathbb{E}_r^N \subseteq \mathbb{E}_2^N$.

The goal of the standardization is to transform a random vector of $\mathbb{E}_2^N$, z, into another, $\tilde{z}$, that has a unit covariance. But if the covariance V of vector z is not invertible, it is necessary to also perform a projection of z onto the range space of V. The standardization procedure proposed here fulfils these two tasks, and may be viewed as a mere PCA.

Let z be a zero-mean random variable of $\mathbb{E}_2^N$, V its covariance matrix and L a matrix such that $V = LL^*$. Matrix L could be defined by any square-root decomposition, such as Cholesky factorization, or a decomposition based on the eigen value decomposition (EVD), for instance. But our preference goes to the EVD because it allows one to handle easily the case of singular or ill-conditioned covariance matrices by projection. More precisely, if

$$V = U \Lambda^2 U^* \tag{2.6}$$

denotes the EVD of V, where Λ is full rank and U possibly rectangular, we define the square root of V as $L = U\Lambda$. Then we can define the standardized variable associated with z:

$$\tilde{z} = \Lambda^{-1} U^* z. \tag{2.7}$$

Note that $(\tilde{z})_i \neq \widetilde{(z_i)}$. In fact, $\widetilde{(z_i)}$ is merely the variable z_i normalized by its variance. In the following, we shall only talk about $\tilde{z}_i$, the ith component of the standardized variable $\tilde{z}$, so avoiding any possible confusion between the similar-looking notations $\tilde{z}_i$ and $\tilde{z}_i$.

Projection and standardization are thus performed all at once if z is not in $\tilde{\mathbb{E}}_2^N$. Algorithm 18 recommends resorting to the singular value decomposition (SVD) instead of EVD in order to avoid the explicit calculation of V and to improve numerical accuracy.

Without restricting the generality we may thus consider in the remaining that the variable observed belongs to $\tilde{\mathbb{E}}_r^N$; we additionally assume that $N > 1$, since otherwise the observation is scalar and the problem does not exist.

2.3. Negentropy, a measure of distance to normality

Define the differential entropy of x as

$$S(p_x) = -\int p_x(u)\log p_x(u)\,du. \tag{2.8}$$

Recall that differential entropy is not the limit of Shannon's entropy defined for discrete variables; it is not invariant by an invertible change of coordinates as the entropy was, but only by orthogonal transforms. Yet, it is the usual practice to still call it entropy, in short. Entropy enjoys very privileged properties as emphasized in [54], and we shall point out in Section 2.3 that information (2.5) may also be written as a difference of entropies.

As a first example, if z is in $\tilde{\mathbb{E}}_2^N$, the entropies of z and $\tilde{z}$ are related by [16]

$$S(p_z) = S(p_{\tilde{z}}) - \tfrac{1}{2}\log\det V. \tag{2.9}$$

If x is zero-mean Gaussian, its pdf will be referred to by the notation $\phi_x(u)$, with

$$\phi_x(u) = (2\pi)^{-N/2}|V|^{-1/2}\exp\{-u^*V^{-1}u\}/2. \tag{2.10}$$

Among the densities of $\tilde{\mathbb{E}}_2^N$ having a given covariance matrix V, the Gaussian density is the one which has the largest entropy. This well-known proposition says that

$$S(\phi_x) \geqslant S(p_y), \quad \forall x, y \in \tilde{\mathbb{E}}_2^N, \tag{2.11}$$

with equality iff $\phi_x(u) = p_y(u)$ almost everywhere. The entropy obtained in the case of equality is

$$S(\phi_x) = \tfrac{1}{2}[N + N\log(2\pi) + \log\det V]. \tag{2.12}$$

For densities in $\tilde{\mathbb{E}}_2^N$, one defines the negentropy as

$$J(p_x) = S(\phi_x) - S(p_x), \tag{2.13}$$

where ϕ_x stands for the Gaussian density with the same mean and variance as p_x. As shown in Appendix A.2, negentropy is always positive, is invariant by any linear invertible change of coordinates, and vanishes if and only if p_x is Gaussian. From (2.5) and (2.13), the mutual information may be written as:

$$I(p_x) = J(p_x) - \sum_{i=1}^{N} J(p_{x_i}) + \frac{1}{2}\log\frac{\prod V_{ii}}{\det V}, \tag{2.14}$$

where V denotes the variance of x (the proof is deferred to Appendix A.3). This relation gives a means of approximating the mutual information, provided we are able to approximate the negentropy about zero, which amounts to expanding the density p_x in the neighbourhood of ϕ_x. This will be the starting point of Section 3.

For the sake of convenience, the transform F defined in Definition 1 will be sought in two steps. In the first step, a transform will cancel the last term of (2.14). It can be shown that this is equivalent to standardizing the data (see Appendix A.4). This step will be performed by resorting only to second-order moments of y. Then in the second step, an orthogonal transform will be computed so that the second term on the right-hand side of (2.14) is minimized while the two others remain constant. In this step, resorting to higher-order cumulants is necessary.

2.4. Measures of statistical dependence

We have shown that both the Gaussian feature and the mutual independence can be characterized with the help of negentropy. Yet, these remarks justify only in part the use of (2.14) as an optimization criterion in our problem. In fact, from Property 2, this criterion should meet the requirements given below.

DEFINITION 5. A contrast is a mapping Ψ from the set of densities $\{p_x, x \in \mathbb{E}^N\}$ to $\mathbb{R}$ satisfying the following three requirements.

- $\Psi(p_x)$ does not change if the components x_i are permuted:

  $\Psi(p_{Px}) = \Psi(p_x), \quad \forall P$ permutation.

- Ψ is invariant by 'scale' change, that is,

  $\Psi(p_{Ax}) = \Psi(p_x), \quad \forall A$ diagonal invertible.

  If x has independent components, then

  $\Psi(p_{Ax}) \leqslant \Psi(p_x), \quad \forall A$ invertible.

DEFINITION 6. A contrast will be said to be discriminating over a set $\mathbb{E}$ if the equality $\Psi(p_{Ax}) = \Psi(p_x)$ holds only when A is of the form ΛP, as defined in Property 2, when x is a random vector of $\mathbb{E}$ having independent components.

Now we are in a position to propose a contrast criterion.

THEOREM 7. The following mapping is a contrast over $\mathbb{E}_2^N$:

$$\Psi(p_z) = -I(p_z).$$

See the appendix for a proof. The criterion proposed in Theorem 7 is consequently admissible for ICA computation. This theoretical criterion, involving a generally unknown density, will be made usable by approximtions in Section 3. Regarding computational loads, the calculation of ICA may still be very heavy even after approximations, and we now turn to a theorem that theoretically explains why the practical algorithm designed in Section 4, that proceeds pairwise, indeed works.

The following two lemmas are utilized in the proof of Theorem 10, and are reproduced below because of their interesting content. Refer to [18, 22] for details regarding Lemmas 8 and 9, respectively.

LEMMA 8 (Lemma of Marcinkiewicz–Dugué (1951)). *The function* $\varphi(u) = e^{P(u)}$, *where* $P(u)$ *is a polynomial of degree* m, *can be a characteristic function only if* $m \leqslant 2$.

LEMMA 9 (Cramér's lemma (1936)). *If* $\{x_i, 1 \leqslant i \leqslant N\}$ *are independent random variables and if* $X = \sum_{i=1}^{N} a_i x_i$ *is Gaussian, then all the variables* x_i *for which* $a_i \neq 0$ *are Gaussian.*

Utilizing these lemmas, it is possible to obtain the following theorem, which can be seen to be a variant of Darmois's theorem (see Appendix A.7).

THEOREM 10. Let x *and* z *be two random vectors such that* $z = Bx$, B *being a given rectangular matrix. Suppose additionally that* x *has independent components and that* z *has pairwise independent components. If* B *has two non-zero entries in the same column* j, *then* x_j *is either Gaussian or deterministic.*

Now we are in a position to state a theorem, from which two important corollaries can be deduced (see Appendices A.7–A.9 for proofs).

THEOREM 11. Let x *be a vector with independent components, of which at most one is Gaussian, and whose densities are not reduced to a point-like mass. Let* C *be an orthogonal* $N \times N$ *matrix and* z *the vector* $z = Cx$. *Then the following three properties are equivalent:*

(i) *The components* z_i *are pairwise independent.*
(ii) *The components* z_i *are mutually independent.*
(iii) $C = \Lambda P$, Λ *diagonal,* P *permutation.*

COROLLARY 12. The contrast in Theorem 7 is discriminating over the set of random vectors having at most one Gaussian component, in the sense of Definition 6.

COROLLARY 13 (Identifiability). *Let no noise be present in model (1.1), and define* $y = Mx$ *and* $y = Fz$, x *being a random variable in some set* $\mathbb{E}$ *of* $\mathbb{E}_2^N$ *satisfying the requirements of Theorem 11. Then if* Ψ *is discriminant over* $\mathbb{E}$, $\Psi(p_z) = \Psi(p_x)$ *if only if* $F = M\Lambda P$, *where* Λ *is an invertible diagonal matrix and* P *a permutation.*

This last corollary is actually stating identifiability conditions of the noiseless problem. It shows in particular that for discriminating contrasts, the indeterminacy is minimal, i.e. is reduced to Property 2; and from Corollary 12, this applies to the contrast in Theorem 7. We recognize in Corollary 13 some statements already emphasized in blind deconvolution problems [21, 52, 64], namely that a non-Gaussian random process can be recovered after convolution by a linear time-invariant stable unknown filter only up to a constant delay and a constant phase shift, which may be seen as the analogues of our indeterminations D and P in Property 2, respectively. For processes with unit variance, we find indeed that $M = FDP$ in the corollary above.

3. Optimization criteria

3.1. Approximation of the mutual information

Suppose that we observe $\tilde{y}$ and that we look for an orthogonal matrix Q maximizing the contrast:

$$\Psi(p_z) = -I(p_z), \quad \text{where} \quad \tilde{z} = Q\tilde{y}. \tag{3.1}$$

In practice, the densities p_z and p_y are not known, so that criterion (3.1) cannot be directly utilized. The aim of this section is to express the contrast in Theorem 7 as a function of the standardized cumulants (of orders 3 and 4), which are quantities more easily accessible. The expression of entropy and negentropy in the scalar case will be first briefly derived. We start with the Edgeworth expansion of type A of a density. A central limit theorem says that if z is a sum of m independent random variables with finite cumulants, then the ith-order cumulant of z is of order

$$\kappa_i \sim m^{(2-i)/2}.$$

This theorem can be traced back to 1928 and is attributed to Cramér [65]. Referring to [39, p. 176, formula 6.49] or to [1], the Edgeworth expansion of the pdf of z up to order 4 about its best Gaussian approximate (here with zero-mean and unit variance) is given by

$$\begin{aligned}\frac{p_z(u)}{\phi_z(u)} = 1 &+ \frac{1}{3!}\kappa_3 h_3(u) \\ &+ \frac{1}{4!}\kappa_4 h_4(u) + \frac{10}{6!}\kappa_3^2 h_6(u) \\ &+ \frac{1}{5!}\kappa_5 h_5(u) + \frac{35}{7!}\kappa_3\kappa_4 h_7(u) \\ &+ \frac{280}{9!}\kappa_3^3 h_9(u) \\ &+ \frac{1}{6!}\kappa_6 h_6(u) + \frac{56}{8!}\kappa_3\kappa_5 h_8(u) + \frac{35}{8!}\kappa_4^2 h_8(u) \\ &+ \frac{2100}{10!}\kappa_3^2\kappa_4 h_{10}(u) + \frac{15\,400}{12!}\kappa_3^4 h_{12}(u) \\ &+ o(m^{-2}).\end{aligned} \tag{3.2}$$

In this expression, κ_i denotes the cumulant of order i of the standardized scalar variable considered (this is the notation of Kendall and Stuart, not assumed subsequently in the multichannel case), and $h_i(u)$ is the Hermite polynomial of degree i, defined by the recursion

$$\begin{aligned}&h_0(u) = 1, \qquad h_1(u) = u, \\ &h_{k+1}(u) = u\,h_k(u) - \frac{\partial}{\partial u}h_k(u).\end{aligned} \tag{3.3}$$

The key advantage of using the Edgeworth expansion instead of Gram–Charlier's lies in the ordering of terms according to their decreasing significance as a function of $m^{-1/2}$. In fact, in the Gram–Charlier expansion, it is impossible to enquire into the magnitude of the various terms. See [39, 40] for general remarks on pdf expansions.

Now let us turn to the expansion of the negentropy defined in (2.13). The negentropy itself can be approximated by an Edgeworth-based expansion as the theorem below shows.

THEOREM 14. For a standardized scalar variable z,

$$\begin{aligned}J(p_z) = &\tfrac{1}{12}\kappa_3^2 + \tfrac{1}{48}\kappa_4^2 + \tfrac{7}{48}\kappa_3^4 \\ &- \tfrac{1}{8}\kappa_3^2\kappa_4 + o(m^{-2}).\end{aligned} \tag{3.4}$$

The proof is sketched in the appendix. Next, from (2.14), the calculus of the mutual information of a standardized vector $\tilde{z}$ needs not only the marginal negentropy of each component z_i but also the joint negentropy of $\tilde{z}$. Now it can be noticed that (see the appendix)

$$J(p_z) = J(p_y), \tag{3.5}$$

since differential entropy is invariant by an orthogonal change of coordinates. Denote by $K_{ij\ldots q}$ the cumulants $\mathrm{Cum}\{z_i, z_j, \ldots, z_q\}$. Then using (3.4) and (3.5), the mutual information of $\tilde{z}$ takes the form, up to $O(m^{-2})$ terms,

$$\begin{aligned}I(p_z) \approx J(p_y) - \\ \frac{1}{48}\sum_{i=1}^{N}\{4K_{iii}^2 + K_{iiii}^2 + 7K_{iii}^4 - 6K_{iii}^2 K_{iii}\}.\end{aligned} \tag{3.6}$$

Yet, cumulants satisfy a multilinearity property [7], which allows them to be called tensors [43, 44]. Denote by Γ the family of cumulants of $\tilde{y}$, and by K those of $\tilde{z} = Q\tilde{y}$. Then this property can be written at orders 3 and 4 as

$$K_{ijk} = \sum_{pqr} Q_{ip}Q_{jq}Q_{kr}\Gamma_{pqr}, \tag{3.7}$$

$$K_{ijkl} = \sum_{pqrs} Q_{ip}Q_{jq}Q_{kr}Q_{ls}\Gamma_{pqrs}. \tag{3.8}$$

On the other hand, $J(p_{\tilde{y}})$ does not depend on Q, so that criterion (3.1) can be reduced up to order m^{-2} to the maximization of a functional ψ:

$$\psi(Q) = \sum_{i=1}^{N} 4K_{iii}^2 + K_{iiii}^2 + 7K_{iii}^4 - 6K_{iii}^2K_{iiii} \tag{3.9}$$

with respect to Q, bearing in mind that the tensors K depend on Q through relations (3.7) and (3.8). Even if the mutual information has been proved to be a contrast, we are not sure that expression (3.9) is a contrast itself, since it is only approximating the mutual information. Other criteria that are contrasts are subsequently proposed, and consist of simplifications of (3.9).

If K_{iii} are large enough, the expansion of $J(p_z)$ can be performed only up to order $O(m^{-3/2})$, yielding $\psi(Q) = 4\sum K_{iii}^2$. On the other hand, if K_{iii} are all null, (3.9) reduces to $\psi(Q) = \sum K_{iiii}^2$. It turns out that these two particular forms of $\psi(Q)$ are contrasts, as shown in the next section. Of course, they can be used even if K_{iii} are neither all large nor all null, but then they would not approximate the contrast in Theorem 7 any more.

3.2. Simpler criteria

The function $\psi(Q)$ is actually a complicated rational function in $N(N-1)/2$ variables, if indices vary in $\{1, \ldots, N\}$. The goal of the remainder of the paper is to avoid exhaustive search and save computational time in the optimization problem, as well as propose a family of contrast functions in Theorem 16, of simpler use. The justification of these criteria is argued and is not subject to the validity of the Edgeworth expansion; in other words, it is not absolutely necessary that they approximate the mutual information.

LEMMA 15. Denote by ${}^rQ$ the matrix obtained by raising each entry of an orthogonal matrix Q to the power r. Then we have

$$\|{}^2Q\boldsymbol{u}\| \leqslant \|\boldsymbol{u}\|.$$

THEOREM 16. The functional

$$\psi(Q) = \sum_{i=1}^{N} K_{ii\ldots i}^2,$$

where $K_{ii\ldots i}$ are marginal standardized cumulants of order r, is a contrast for any $r \geqslant 2$. Moreover, for any $r \geqslant 3$, this contrast is discriminating over the set of random vectors having at most one null marginal cumulant of order r. Recall that the cumulants are functions of Q via the multilinearity relation, e.g. (3.7) and (3.8).

The proofs are given in the appendix (see also [15] for complementary details). These contrast functions are generally less discriminating than (3.1) (i.e. discriminating over a smaller subset of random vectors). In fact, if two components have a zero cumulant of order r, the contrast in Theorem 16 fails to separate them (this is the same behaviour as for Gaussian components). However, as in Theorem 11, at most one source component is allowed to have a null cumulant.

Now, it is pertinent to notice that the quantity

$$\Omega_r = \sum_{i_1\ldots i_r} K_{i_1 i_2 \ldots i_r}^2 \tag{3.10}$$

is invariant under linear and invertible transformations (cf. the appendix). This result gives, in fact, an important practical significance to contrast functions such as in Theorem 16. Indeed, the maximization of $\psi(Q)$ is equivalent to the minimization of $\Omega - \psi(Q)$, which is eventually the same as minimizing the sum of the squares of all cross-cumulants of order r, and these cumulants are precisely the measure of statistical dependence at order r. Another consequence of this invariance is that if $\Psi(y) = \Psi(x)$, where x has uncorrelated components at orders involved in the definition of ψ, then y also has independent components in the same sense. A similar interpretation can be given for

expression (3.9) since

$$\Omega_{3,4} = \sum_{ijklmnqr} 4K_{ijk}^2 + K_{ijkl}^2 + 3K_{ijk}K_{ijn}K_{kqr}K_{qrn} + 4K_{ijk}K_{imn}K_{jmr}K_{knr} - 6K_{ijk}K_{ilm}K_{jklm} \tag{3.11}$$

is also invariant under linear and invertible transformations [16] (see the appendix for a proof). However, on the other hand, the minimization of (3.11) does not imply individual minimization of cross-cumulants any more, due to the presence of a minus sign, among others.

The interest in maximizing the contrast in Theorem 16 rather than minimizing the cross-cumulants lies essentially in the fact that only N cumulants are involved instead of $O(N^r)$. Thus, this spares us a lot of computations when estimating the cumulants from the data. A first analysis of complexity was given in [15], and will be addressed in Section 4.

3.3. Link with blind identification and deconvolution

Criterion (3.9) and that in Theorem 16 may be connected with other criteria recently proposed in the literature for the blind deconvolution problem [3, 5, 21, 31, 48, 52, 64]. For instance, the criterion proposed in [52] may be seen to be equivalent to maximize $\sum K_{iiii}^2$. In [21], one of the optimization criteria proposed amounts to minimizing $\sum S(p_{z_i})$, which is consistent with (2.13), (2.14) and (3.1). In [5], the family of criteria proposed contains (3.1).

On the other hand, identification techniques presented in [35, 45, 46, 57, 63] solve a system of equations obtained by equation error matching. Although they work quite well in general, these approaches may seem arbitrary in their selection of particular equations rather than others. Moreover, their robustness is questioned in the presence of measurement noise, especially non-Gaussian, as well as for short data records. In [62] a matching in the least-squares (LS) sense is proposed; in [12] the use of much more equations than unknowns improves on the robustness for short data records. A more general approach is developed in [28], where a weighted LS matching is suggested. Our equation (3.9) gives a possible justification to the process of selecting particular cumulants, by showing their dominance over the others. In this context, some simplifications would occur when developing (3.9) as a function of Q since the components y_i are generally assumed to be identically distributed (this is not assumed in our derivations). However, the truncation in the expansion has removed the optimality character of criterion (3.1), whereas in the recent paper [34] asymptotic optimality is addressed.

4. A practical algorithm

4.1. Pairwise processing

As suggested by Theorem 11, in order to maximize (3.9) it is necessary and sufficient to consider only pairwise cumulants of $\tilde{z}$. The proof of Theorem 11 did not involve the contrast expression, but was valid only in the case where the observation y was effectively stemming linearly from a random variable x with independent components (model (1.1) in the noiseless case). It turns out that it is true for any $\tilde{y}$ and any $\tilde{z} = Q\tilde{y}$, and for any contrast of polynomial (and by extension, analytic) form in marginal cumulants of $\tilde{z}$ as is the case in (3.9) or Theorem 16.

THEOREM 17. Let $\psi(Q)$ be a polynomial in marginal cumulants denoted by K. Then the equation $d\psi = 0$ imposes conditions only on components of K having two distinct indices. The same statement holds true for the condition $d^2\psi < 0$.

The proof resorts to differentiation tools borrowed from classical analysis, and not to statistical considerations (see the appendix). The theorem says that pairwise independence is sufficient, provided a contrast of polynomial form is utilized. This motivated the algorithm developed in this section.

Given an $N \times T$ data matrix, $Y = \{y(t), 1 \leqslant t \leqslant T\}$, the proposed algorithm processes each pair in turn (similarly to the Jacobi algorithm in the diagonalization of symmetric real matrices); $Z_{i:}$ will denote the ith row of a matrix Z.

ALGORITHM 18

(1) Compute the SVD of the data matrix as $Y = V\Sigma U^*$, where V is $N \times \rho$, Σ is $\rho \times \rho$ and U^* is $\rho \times T$, and ρ denotes the rank of Y. Use therefore the algorithm proposed in [11] since T is often much larger than N. Then $Z = \sqrt{T}U^*$ and $L = V\Sigma/\sqrt{T}$.
(2) Initialize $F = L$.
(3) Begin a loop on the sweeps: $k = 1, 2, \ldots, k_{\max}$; $k_{\max} \leqslant 1 + \sqrt{\rho}$.
(4) Sweep the $\rho(\rho - 1)/2$ pairs (i, j), according to a fixed ordering. For each pair:
  (a) estimate the required cumulants of $(Z_{i:}, Z_{j:})$, by resorting to κ-statistics for instance [39, 44],
  (b) find the angle α maximizing $\psi(Q^{(i,j)})$, where $Q^{(i,j)}$ is the plane rotation of angle α, $\alpha \in]-\pi/4, \pi/4]$, in the plane defined by components $\{i, j\}$,
  (c) accumulate $F := FQ^{(i,j)*}$,
  (c) update $Z := Q^{(i,j)}Z$.
(5) End of the loop on k if $k = k_{\max}$, or if all estimated angles are very small (compared to $1/T$).
(6) Compute the norm of the columns of F: $\Delta_{ii} = \| F_{:i} \|$.
(7) Sort the entries of Δ in decreasing order: $\Delta := P\Delta P^*$ and $F := FP^*$.
(8) Normalize F by the transform $F := F\Delta^{-1}$.
(9) Fix the phase (sign) of each column of F according to Definition 4(e). This yields $F := FD$.

The core of the algorithm is actually step 4(b), and as shown in [15] it can be carried out in various ways. This step becomes particularly simple when a contrast of the type given by Theorem 16 is used. Let us take for example the contrast in Theorem 16 at order 4 (a similar and less complicated reasoning can be carried out at order 3).

Step 4(b) of the algorithm maximizes the contrast in dimension 2 with respect to orthogonal transformations. Denote by Q the Givens rotation

$$Q = \frac{1}{\sqrt{1+\theta^2}} \begin{bmatrix} 1 & \theta \\ -\theta & 1 \end{bmatrix}$$

and $\xi = \theta - 1/\theta$. (4.1)

Then the contrast takes the same value at θ and $-1/\theta$. Yet, it is a rational function of θ, so that it can be expressed as a function of variable ξ only. More precisely, we have

$$\psi(\xi) = [\xi^2 + 4]^{-2} \sum_{k=0}^{4} b_k \xi^k, \tag{4.2}$$

where coefficients b_k are given in the appendix. Here, we have somewhat abused the notations, and write ψ as a function of ξ instead of Q itself. However, there is no ambiguity since θ characterizes Q completely, and ξ determines two equivalent solutions in θ via the equation

$$\theta^2 - \xi\theta - 1 = 0. \tag{4.3}$$

Expression (4.2) extends to the case of complex observations, as demonstrated in the appendix. The maximization of (4.2) with respect to variable ξ leads to a polynomial rooting which does not raise any difficulty, since it is of low degree (there even exists an analytical solution since the degree is strictly smaller than 5):

$$\omega(\xi) = \sum_{k=0}^{4} c_k \xi^k = 0. \tag{4.4}$$

The exact value of coefficients c_k is also given in the appendix. Thus, step 4(b) of the algorithm consists simply of the following four steps:

- compute the roots of $\omega(\xi)$ by using (A.39);
- compute the corresponding values of $\psi(\xi)$ by using (A.38);
- retain the absolute maximum; (4.5)
- root Eq. (4.3) in order to get the solution θ_0 in the interval $]-1, 1]$.

The solution θ_0 corresponds to the tangent of the rotation angle. Since it is sufficient to have one representative of the equivalence class in Property 2, the solution in $]-1, 1]$ suffices. Obvious real-time implementations are described in [14]. Additional details on this algorithm may also be found in [15]. In particular, in the noiseless case the polynomial (4.4) degenerates and we end up with the rooting of a polynomial of degree 2 only. However, it is wiser to resort to the latter simpler algorithm only when $N = 2$ [15].

4.2. Convergence

It is easy to show [15] that Algorithm 18 is such that the global contrast function $\psi(Q)$ monotonically increases as more and more iterations are run. Since this real positive sequence is bounded above, it must converge to a maximum. How many sweeps ($k_{\max}$) should be run to reach this maximum depends on the data. The value of $k_{\max}$ has been chosen on the basis of extensive simulations: the value $k_{\max} = 1 + \sqrt{N}$ seems to give satisfactory results. In fact, when the number of sweeps, k, reaches this value, it has been observed that the angles of all plane rotations within the last sweep were very small and that the contrast function had reached its stationary value. However, iterations can obviously be stopped earlier.

In [52], the deconvolution algorithm proposed has been shown to suffer from spurious maxima in the noisy case. Moreover, as pointed out in [64], even in the noiseless case, spurious maxima may exist for data records of limited length. This phenomenon has not been observed when running ICA, but could a priori also happen. In fact, nothing ensures the absence of local maxima, at least based on the results we have presented up to now. Algorithm 18 finds a sequence of absolute maxima with respect to each plane rotation, but no argument has been found that ensures reaching the global absolute maximum with respect to the cumulated rotation. So it is only conjectured that this search is sufficient over the group of orthogonal matrices.

4.3. Computational complexity

When evaluating the complexity of Algorithm 18, we shall count the number of floating point operations (flops). A flop corresponds to a multiplication followed by an addition. Note that the complexity is often assimilated to the number of multiplications, since most of the time there are more multiplications than additions. In order to make the analysis easier, we shall assume that N and T are large, and only retain the terms of highest orders.

In step 1, the SVD of a matrix of size $N \times T$ is to be calculated. Since T will necessarily be larger than N, it is appropriate to resort to a special technique proposed by Chan [11] consisting of running a QR factorization first. The complexity is then of $3TN^2 - 2N^3/3$. Making explicit the matrix L requires the multiplication of an $N \times \rho$ matrix by a diagonal matrix of size ρ. This requires an additional $N\rho$ flops, which is negligible with respect to TN^2.

In step 4(a), the five pairwise cumulants of order 4 must be calculated. If $z_{i:}$ and $z_{j:}$ are the two row vectors concerned, this may be done as follows for reasonable values of T ($T > 100$):

$$
\begin{array}{lll}
\zeta_{ii} = z_{i:} \cdot z_{i:}, & \text{termwise product} & T \text{ flops} \\
\zeta_{jj} = z_{j:} \cdot z_{j:}, & \text{termwise product} & T \text{ flops} \\
\zeta_{ij} = z_{i:} \cdot z_{j:}, & \text{termwise product} & T \text{ flops} \\
\Gamma_{iiii} = \zeta_{ii}^* \zeta_{ii}/T - 3, & \text{scalar product} & T+1 \text{ flops} \\
\Gamma_{iiij} = \zeta_{ii}^* \zeta_{ij}/T, & \text{scalar product} & T+1 \text{ flops} \\
\Gamma_{iijj} = \zeta_{ii}^* \zeta_{jj}/T - 1, & \text{scalar product} & T+1 \text{ flops} \\
\Gamma_{ijjj} = \zeta_{ij}^* \zeta_{jj}/T, & \text{scalar product} & T+1 \text{ flops} \\
\Gamma_{jjjj} = \zeta_{jj}^* \zeta_{jj}/T - 3, & \text{scalar product} & T+1 \text{ flops}
\end{array}
\tag{4.6}
$$

So step 4(a) requires $O(8T)$ flops. The polynomial rooting of step 4(b) requires $O(1)$ flops, as well as the selection of the absolute maximum. The accumulation of Step 4(c) needs $4N$ flops because only two columns of F are changed. The update Step 4(d) is eventually calculated within $4T$ flops.

If the loop on k is executed K times, then the complexity above is repeated $O(K\rho^2/2)$ times. Since $\rho \leqslant N$, we have a complexity bounded by $(12T + 4N)KN^2/2$. The remaining normalization steps require a global amount of $O(N\rho)$ flops.

As conclusion, and if $T \gg N$, the execution of Algorithm 18 has a complexity of

$$O(6KN^2T) \text{ flops.} \tag{4.7}$$

In order to compare this with usual algorithms in numerical analysis, let us take realistic examples. If $T = O(N^2)$ and $K = O(N^{1.2})$, the global complexity is $O(N^{9.2})$, and if $T = O(N^{3.2})$, we get $O(N^4)$ flops. These orders of magnitude may be compared

to usual complexities in $O(N^3)$ of most matrix factorizations.

The other algorithm at our disposal to compute the ICA within a polynomial time is the one proposed by Cardoso (refer to [10] and the references therein). The fastest version requires $O(N^5)$ flops, if all cumulants of the observations are available and if N sources are present with a high signal to noise ratio. On the other hand, there are $O(N^4/24)$ different cumulants in this 'supersymmetric' tensor. Since estimating one cumulant needs $O(\alpha T)$ flops, $1 < \alpha \leqslant 2$, we have to take into account an additional complexity of $O(\alpha N^4 T/24)$. In this operator-based approach, the computation of the cumulant tensor is the task that is the most computationally heavy. For $T = O(N^2)$ for instance, the global complexity of this approach is roughly $O(N^6)$. Even for $T = O(N)$, which is too small, we get a complexity of $O(N^5)$. Note that this is always larger than (4.9). The reason for this is that only pairwise cumulants are utilized in our algorithm.

Lastly, one could think of using the multilinearity relation (3.8) to calculate the five cumulants required in step 4(a). This indeed requires little computational means, but needs all input cumulants to be known, exactly as in Cardoso's method above. The computational burden is then dominated by the computation of input cumulants and represents $O(TN^4)$ flops.

5. Simulation results

5.1. Performance criterion

In this section, the behaviour of ICA in the presence of non-Gaussian noise is investigated by means of simulations. We first need to define a distance between matrices modulo a multiplicative factor of the form ΛP, where Λ is diagonal invertible and P is a permutation. Let A and $\hat{A}$ be two invertible matrices, and define the matrices with unit-norm columns

$$\underline{A} = A\Delta^{-1}, \quad \underline{\hat{A}} = \hat{A}\hat{\Delta}^{-1},$$

$$\text{with } \Delta_{kk} = \|A_{:k}\|, \quad \hat{\Delta}_{kk} = \|\hat{A}_{:k}\|.$$

The gap $\varepsilon(A, \hat{A})$ is built from the matrix $D = \underline{A}^{-1}\underline{\hat{A}}$ as

$$\varepsilon(A, \hat{A}) = \sum_i \left| \sum_j |D_{ij}| - 1 \right|^2 + \sum_j \left| \sum_i |D_{ij}| - 1 \right|^2 + \sum_i \left| \sum_j |D_{ij}|^2 - 1 \right| + \sum_j \left| \sum_i |D_{ij}|^2 - 1 \right|. \tag{5.1}$$

It is shown in the appendix that this measure of distance is indeed invariant by postmultiplication by a matrix of the form ΛP:

$$\varepsilon(A\Lambda P, \hat{A}) = \varepsilon(A, \hat{A}\Lambda P) = \varepsilon(A, \hat{A}). \tag{5.2}$$

Moreover, $\varepsilon(A, \hat{A}) = 0$ if and only if A and $\hat{A}$ can be deduced from each other by multiplication by a factor ΛP:

$$\varepsilon(A, \hat{A}) = 0 \Leftrightarrow \hat{A} = A\Lambda P. \tag{5.3}$$

5.2. Behaviour in the presence of non-Gaussian noise when $N = 2$

Consider the observations in dimension N for realization t:

$$y(t) = (1 - \mu)Mx(t) + \mu\beta w(t), \tag{5.4}$$

where $1 \leqslant t \leqslant T$, x and w are zero-mean standardized random variables uniformly distributed in $[-\sqrt{3}, \sqrt{3}]$ (and thus have a kurtosis of -1.2), μ is a positive real parameter of $[0, 1]$ and $\beta = \|M\|/\|I\|$, $\|\cdot\|$ denoting the L^2 matrix norm (the largest singular value). Then when μ ranges from 0 to 1, the signal to noise kurtosis ratio decreases from infinity to zero. Since x and w have the same statistics, the signal to noise ratio can be indifferently defined at orders 2 or 4. One may assume the following as a definition of the signal to noise ratio (see Fig. 1):

$$\mathrm{SNR}_{\mathrm{dB}} = 10 \log_{10} \frac{1 - \mu}{\mu}.$$

In this section we assume $N = 2$ and

$$M = \begin{bmatrix} 1 & 3 \\ -3 & 1 \end{bmatrix}. \tag{5.5}$$

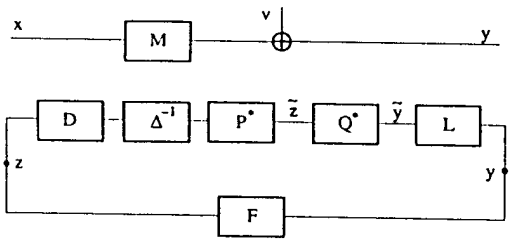

Fig. 1. Identification block diagram. We have in particular $y = Fz = LQ^*(P^*\Delta^{-1}D)z$. The matrix $(P^*\Delta^{-1}D)$ is optional and serves to determine uniquely a representative of the equivalence class of solutions.

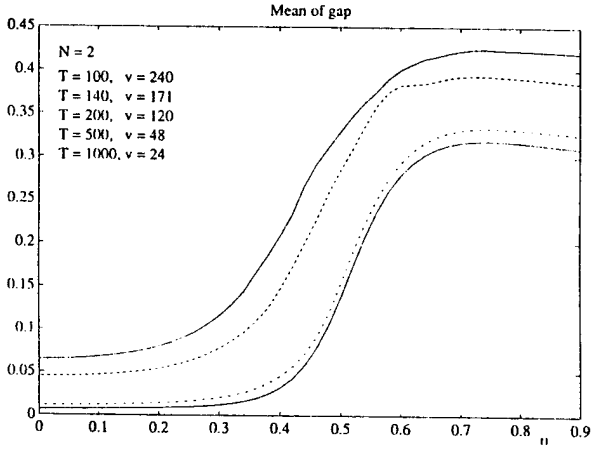

Fig. 2. Average values of the gap ε between the true mixing matrix M and the estimated one, F, as a function of the noise rate and for various data lengths. Here the dimension is $N = 2$, and averages are computed over v trials.

We represent in Figs. 2–4 the behaviour of Algorithm 18 when the contrast in Theorem 16 is used at order $r = 4$ (which also coincides with (3.9) since the densities are symmetrically distributed in this simulation). The mean of the error ε defined in (5.1) has been plotted as a function of μ in Fig. 2, and as a function of T in Fig. 3. The mean was calculated over a number v of averagings (indicated in Fig. 2) chosen so that the product vT is constant: $vT \approx 24000$. The curves presented in Fig. 2 have been calculated for μ ranging from 0 to 1 by steps of 0.02. Fig. 4 shows the standard deviations of ε obtained with the same data. A similar simulation was presented in [16], where signal and noise had different kurtosis.

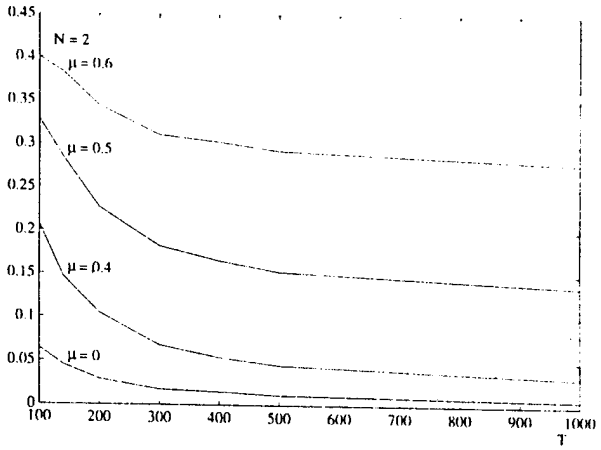

Fig. 3. Average values of the gap ε as a function of the data length T and for various noise rates μ. The dimension is $N = 2$.

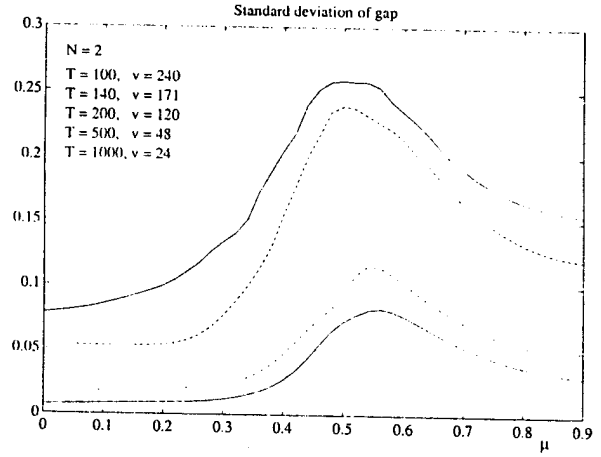

Fig. 4. Standard deviations of the gap ε for the same input and output data as for Fig. 2.

Most surprising in these simulations in the excellent behaviour of ICA even in the close neighbourhood of $\mu = 0.5$ (viz. 0 dB); keep in mind that the μ-scale is zooming about 0 dB (according to Fig. 9, an SNR of 1 dB corresponds to $\mu = 0.44$ for example). The exact value of μ for which the ICA remains interesting depends on the integration length T. In fact, the shorter the T, the larger the apparent noise, since estimation errors are also seen by the algorithm as extraneous noise. In Fig. 3, for instance, it can be noticed that the calculation of ICA with $T = 100$ in the noiseless case is about as good as with $T = 300$ when $\mu = 0.4$. For larger

values of μ, the ICA algorithm considers the vector Mx as a noise and the vector $\mu\beta w$ as a source vector. Because w has independent components, it could be checked out conversely that $\varepsilon(I, F)$ tends to zero as μ tends to one, showing that matrix F is approaching AP.

5.3. *Behaviour in the noiseless case when $N > 2$*

Another interesting experiment is to look at the influence of the choice of the sweeping strategy on the convergence of the algorithm. For this purpose, consider now 10 independent sources. In Algorithm 18, the pairs are described according to a standard cyclic-by-rows ordering, but any other ordering may be utilized. To see the influence of this ordering without modifying the code, we have changed the order of the source kurtosis. In ordering 1, the source kurtosis are

$$[1 \quad -1 \quad 1 \quad -1 \quad 1.5 \quad -1.5 \quad 2 \quad -2 \quad 1 \quad -1] \tag{5.6}$$

whereas in ordering 2 and 3 they are, respectively,

$$[1 \quad 2 \quad 1.5 \quad 1 \quad 1 \quad -1 \quad -1.5 \quad -2 \quad -1 \quad -1]$$

and (5.7)

$$[1 \quad 1 \quad 1.5 \quad 2 \quad 1 \quad -1 \quad -1.5 \quad -1 \quad -1 \quad -2].$$

The presence of cumulants of opposite signs does not raise any obstacle, as well as the presence of at most one null cumulant as shown in [16]. The mixing matrix, M, is Toeplitz circulant with the vector

$$[3 \quad 0 \quad 2 \quad 1 \quad -1 \quad 1 \quad 0 \quad 1 \quad -1 \quad -2] \tag{5.8}$$

as first row.

No noise is present. This simulation is performed directly from cumulants, using unavoidably the procedure described in the last paragraph of Section 4.3, so that the performances are those that would be obtained for $T = \infty$ with real-world signals (see Appendix A.21). The contrast utilized is that in Theorem 16 with $r = 4$. The evolution of gap ε between the original matrix, M, and the estimated matrix, F, is plotted in Fig. 5 for the three orderings, as a function of the number of Givens rotations computed. Observe that the gap ε is not necessarily monotonically decreasing, whereas the contrast always is, by construction of the algorithm. For convenience, the contrast is also represented in Fig. 6. The contrast reaches a stationary value of 18.5 around the 90th iteration, that is at the end of the second sweep. It can be noticed that this corresponds precisely to the sum of the squares of source cumulants (5.6), showing that the upper bound is indeed reached. Readers desiring to program the ICA should pay attention to the fact that

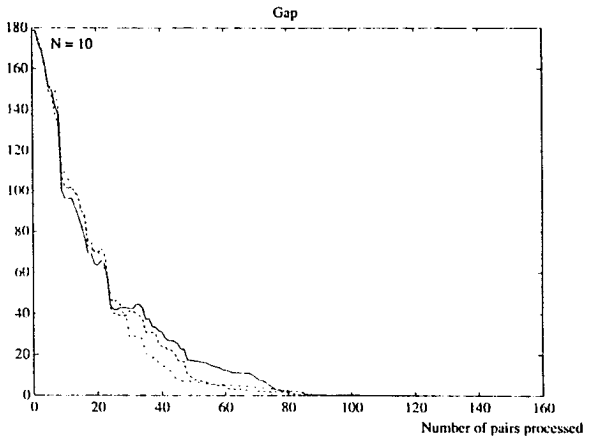

Fig. 5. Gap between the true mixing matrix M and the estimated one, F, as a function of the number of iterations. These are asymptotic results, i.e. the data length may be considered infinite. Solid and dashed lines correspond to three different orderings. The dimension is $N = 10$.

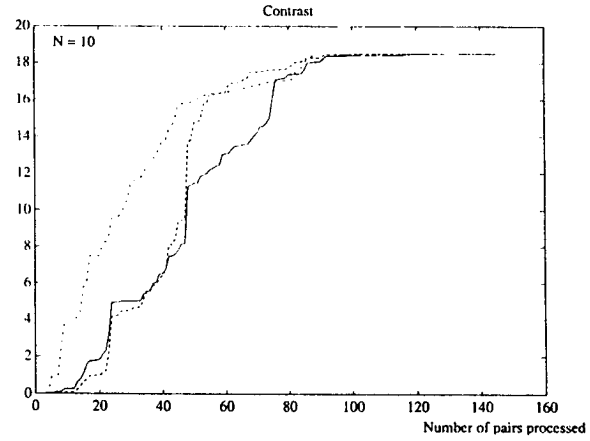

Fig. 6. Evolution of the contrast function as more and more iterations are performed, in the same problem as shown in Fig. 5, with the same three orderings.

the speed of convergence can depend on the standardization code (e.g. changing the sign of a singular vector can affect the behaviour of the convergence).

It is clear that the values of source kurtosis can have an influence on the convergence speed, as in the Jacobi algorithm (provided they are not all equal of course). In the latter algorithm, it is well known that the convergence can be speeded up by processing the pair for which non-diagonal terms are the largest. Here, the difference is that cross-terms are not explicitly given. Consequently, even if this strategy could also be applied to the diagonalization of the tensor cumulant, the computation of all pairwise cross-cumulants at each iteration is necessary. We would eventually loose more in complexity than we would gain in convergence speed.

5.4. *Behaviour in the presence of noise when $N > 2$*

Now, it is interesting to perform the same experiment as in Section 5.2, but for values of N larger than 2. We again take the experiment described by (5.4). The components of x and w are again independent random variables identically and uniformly distributed in $[-\sqrt{3}, \sqrt{3}]$, as in Section 5.2. In Fig. 7, we report the gap ε that we obtain for $N = 10$. The matrix M was in this example again the Toeplitz circulant matrix of Section 5.3 with a first row defined by

$$[3\ 0\ 2\ 1\ -1\ 1\ 0\ 1\ -1\ -2].$$

It may be observed in Fig. 7 that an integration length of $T = 100$ is insufficient when $N = 10$ (even in the noiseless case), whereas $T = 500$ is acceptable. For $T = 100$, the gap ε showed a very high variance, so that the top solid curve is subject to important errors (see the standard deviations in Fig. 8); the number ν of trials was indeed only 50 for reasons of computational complexity. As in the case of $N = 2$ and for larger values of T, the increase of ε is quite sudden when μ increases. With a logarithmic scale (in dB), the sudden increase would be even much steeper. After inspection, the ICA can be computed with Algorithm 18 up to signal to noise ratios of about $+2$ dB ($\mu = 0.35$)

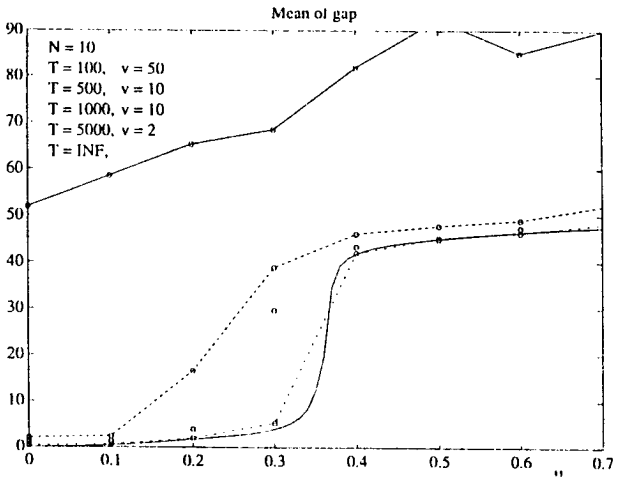

Fig. 7. Averages of the gap ε between the true mixing matrix M and the estimated one, F, as a function of the noise rate and for various data lengths, T. Here the dimension is $N = 10$. Notice the gap is small below a 2 dB kurtosis signal to noise ratio (rate $\mu = 0.35$) for a sufficient data length. The bottom solid curve corresponds to ultimate performances with infinite integration length.

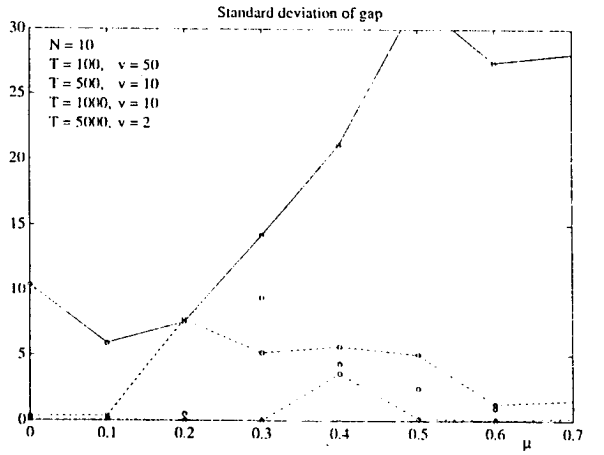

Fig. 8. Standard deviations of the gap ε obtained for integration length T and computed over ν trials, for the same data as for Fig. 7.

for integration lengths of the order of 1000 samples. The continuous bottom solid curve in Fig. 7 shows the performances that would be obtained for infinite integration length. In order to compute the

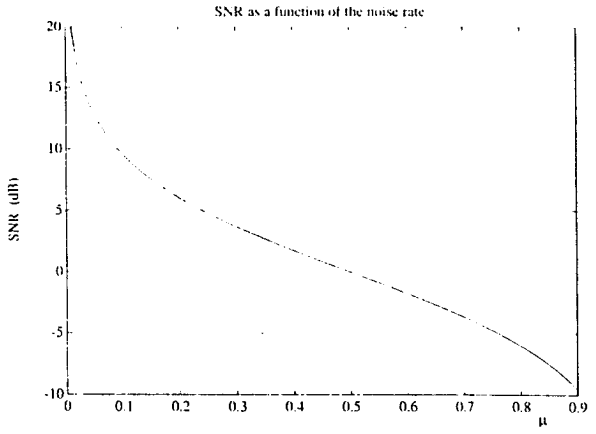

Fig. 9. SNR as a function of the noise rate μ.

ICA in such a case, the *validation algorithm* described in Appendix A.21 has been utilized, as in Section 5.3. It can be inferred from this curve that, for $N = 10$, an intregration of $T = 5000$ reaches performances very close to the ultimate.

Source MATLAB codes of these experiments can be obtained from the author upon request. Other simulation results related to ICA are also reported in [15].

6. Conclusions

The definition of ICA given within the framework of this paper depends on a contrast function that serves as a rotation selection criterion. One of the contrasts proposed is built from the mutual information of standardized observations. For practical purposes this contrast is approximated by the Edgeworth expansion of the mutual information, and consists of a combination of third- and fourth-order marginal cumulants.

A particular contrast of interest is the sum of squares of marginal cumulants of order 4. An efficient algorithm is proposed that maximizes this criterion, by rooting a sequence of polynomials of degree 4. The overall complexity of the algorithm depends on the integration length, but it is reasonable to say that N^4 to N^5 floating point operations are required to compute an ICA in dimension N.

Simulation results demonstrate convergence and robustness of the algorithm, even in the presence of non-Gaussian noise (up to 0 dB for a noise having the same statistics as the sources), and with limited integration lengths. The influence of finite integration and non-Gaussian noise are recent investigations in the area of high-order statistics.

It is envisaged that this tool might be useful in antenna array processing (separation and recognition of sources from unknown arrays, localization from perturbed arrays, jammers rejection, etc.), Bayesian classification, data compression, as well as many areas where PCA has been already utilized for many years, such as detection, linear whitening or exploratory analysis. However, it has not been proved (but also never been observed to happen) that the algorithm proposed cannot be stuck at some local maximum. The algorithm is indeed guaranteed to reach the global maximum only in the case of two sources in the presence of non-Gaussian noise.

7. Appendix A

7.A.1. Proof of relation (1.2)

Since x is non-degenerate, it admits an invertible $\rho \times \rho$ covariance matrix, which we may denote by Δ^{-2} for convenience, where matrix Δ is diagonal, invertible and positive real. Denote by an asterisk transposition and complex conjugation. The covariance of z can be written as $C\Delta^{-2}C^*$, and must be diagonal since z has independent components. Thus, denote by Λ^2 this covariance matrix, where Λ is diagonal and positive real (since C is full column rank, Λ has exactly $N - \rho$ null entries). Thus, we have $C\Delta^{-2}C^* = \Lambda^2$. Let P be a permutation matrix such that the ρ first diagonal entries of $\Lambda' = P\Lambda P^*$ are non-zero. Denote by Λ_ρ the $\rho \times \rho$ block formed of these entries. Then the relation $(PC\Delta^{-1})(PC\Delta^{-1})^* = \Lambda'^2$ shows that the $N - \rho$ rows of matrix $A = PC\Delta^{-1}$ are null. Denote by A_ρ the block formed of the ρ first rows. Then there exists a unitary matrix Q_ρ such that $A_\rho = \Lambda_\rho Q_\rho$. This is equivalent to

$$A = \begin{bmatrix} \Lambda_\rho Q_\rho \\ 0 \end{bmatrix} = \Lambda' \begin{bmatrix} Q_\rho \\ 0 \end{bmatrix}.$$

Denote by Q' the last $N \times \rho$ block matrix. By construction, we consequently have $C = P^* \Lambda' Q' \Delta$, which can be written as $C = \Lambda Q \Delta$, with $Q = P^* Q'$ and $Q'^* Q' = I$. □

7.A.2. Properties of (2.13)

Adding and subtracting a term in the definition of negentropy gives

$$J(p_x) = \int \log p_x(u) p_x(u)\,du - \int \log \phi_x(u) p_x(u)\,du + \int \log \phi_x(u) p_x(u)\,du - \int \log \phi_x(u) \phi_x(u)(u)\,du,$$

which can be written as

$$J(p_x) = \int \log \frac{p_x}{\phi_x} p_x(u)\,du + \int \log \phi_x(u)[p_x(u) - \phi_x(u)]\,du. \tag{A.1}$$

Now, by definition, $\phi_x(u)$ and $p_x(u)$ have the same first- and second-order moments. Yet, since $\log \phi_x(u)$ is a polynomial of degree 2, we have

$$\int \log \phi_x(u) \phi_x(u)\,du = \int \log \phi_x(u) p_x(u)\,du. \tag{A.2}$$

Combining the two last equations shows that negentropy (2.13) may be written in another manner, as a Kullback divergence:

$$J(p_x) = \int p_x(u) \log \frac{p_x(u)}{\phi_x(u)}\,du. \tag{A.3}$$

This proves, referring to (2.4), that

$$J(p_x) \geqslant 0, \tag{A.4}$$

with equality iff $\phi_x = p_x$ almost everywhere. Again as a Kullback divergence between two densities defined on the same space, negentropy is invariant by an invertible change of coordinates.

7.A.3. Proof of (2.14)

From (2.13) we start with

$$J(p_x) - \sum_i J(p_{x_i}) = S(\phi_x) - S(p_x) - \sum_i S(\phi_{x_i}) + \sum_i S(p_{x_i}).$$

Using now (2.5),

$$J(p_x) - \sum_i J(p_{x_i}) = I(p_x) + S(\phi_x) - \sum_i S(\phi_{x_i}). \tag{A.5}$$

Eq. (2.14) is then obtained by replacing Gaussian entropies by their values given in (2.12). □

7.A.4. Necessary and sufficient condition for the last term of (2.14) to be zero

The last term of (2.14) is zero when

$$\det V = \prod_{i=1}^{N} V_{ii}. \tag{A.6}$$

It is obvious that if V is diagonal, (A.6) is satisfied. Let us prove the reverse, and assume that (A.6) holds. Then write the Cholesky factorization of covariance V as $V = LL^*$. This yields

$$V_{ii} = \sum_{k=1}^{i} L_{ik}^2 \tag{A.7}$$

and after substitution in (A.6)

$$\prod_{i=1}^{N} L_{ii}^2 = \prod_{i=1}^{N} \left(L_{ii}^2 + \sum_{k=1}^{i-1} L_{ik}^2 \right). \tag{A.8}$$

This relation is possible only if either some row of L is null or when all L_{ik} are null. In other words, if V is invertible, it must be diagonal. □

7.A.5. Proof of Theorem 7

The second requirement of Definition 5 is satisfied because the random variable is standardized. Regarding the third one, note first that from (2.4) or (2.5) the mutual information can cancel only when

all components are mutually independent. Again because of (2.4), the information is always positive. Thus, we indeed have $\Psi(Ax) \leqslant 0$, with equality if and only if the components are independent.

At this stage a comment is relevant. In scalar blind deconvolution problems [21], in order to prove that the minimization of entropy was an acceptable criterion, it was necessary to resort to the property

$$S(\sum a_i x_i) - S(x) \geqslant \log(\sum a_i^2)/2, \qquad \text{(A.9)}$$

satisfied as soon as the random variables x_i are independently and identically distributed (i.i.d.) according to the same distribution as x. Directions for the proof of (A.9) may be found in [6]. In our framework, the proof was somewhat simpler and property (A.9) was not necessary. This stems from the fact that we are dealing with multivariate random variables, and thus multivariate entropy. In return, the components of x are not requested to be i.i.d. □

7.A.6. Proofs of Lemma 8 and 9

Lemma 8 was first proved by Marcinkiewicz in 1940, but Dugué gave a shorter proof in 1951 [22]. Extensions of these lemmas may be found in Kagan et al. [38], but have the inconvenience of being more difficult to prove. Lemma 9 is due to Cramér and may be found in his book of 1937 [18]. See also [38, p. 21] for extensions. Some general comments on these lemmas can also be found in [19]. □

7.A.7. Remarks on Theorem 10

This theorem may be seen to be a direct consequence of a theorem due to Darmois [19]. This was unnoticed in [15].

THEOREM 19 (Darmois' theorem (1953)). *Define the two random variables X_1 and X_2 as*

$$X_1 = \sum_{i=1}^{N} a_i x_i, \qquad X_2 = \sum_{i=1}^{N} b_i x_i,$$

where x_i are independent random variables. Then if X_1 and X_2 are independent, all variables x_j for which $a_j b_j \neq 0$ are Gaussian.

This theorem can also be found in [19; 49, p. 218; 38, p. 89]. Its proof needs Lemmas 8 and 9, and also the following lemma published in 1953 by Darmois [19].

LEMMA 20 (Darmois' lemma). *Let f_1, f_2, g_1 and g_2 be continuous functions in an open set U, and null at the origin. If*

$$f_1(au + cv) + f_2(bu + dv) = g_1(u) + g_2(v),$$

$$\forall u, v \in U,$$

where a, b, c are non-zero constants such that $ad \neq bc$, then $f_i(x)$ are polynomials of degree $\leqslant 2$.

The proof of the lemma makes use of successive finite differences [38, 19, p. 5]. More general results of this kind can be found in Kagan et al. [38].

7.A.8. Proof of Theorem 11

Implications (iii) ⇒ (ii) and (ii) ⇒ (i) are quite obvious. We shall prove the last one, namely (i) ⇒ (iii). Assume z has pairwise independent components, and suppose C is not of the form ΛP. Since C is orthogonal, it necessarily has two non-zero entries in at least two different columns. Then by applying Theorem 10 twice, x has at least two Gaussian components, which is contrary to our hypothesis. □

7.A.9. Proofs of Corollaries 12 and 13

To prove Corollary 12, assume that $\Psi(p_{Ax}) = \Psi(p_x)$, where x has independent components. Then $\Psi(p_x) = 0$ because of (2.4), which yields $\Psi(p_{Ax}) = 0$. Next, again because of (2.4), Ax has independent components. Now Theorem 11 implies that $A = \Lambda P$, which was to be proved. On the other hand, the reverse implication is immediate since $\Psi(p_{\Lambda P x}) = \Psi(p_x)$ is satisfied by any contrast; see Definition 5 and Theorem 7.

To prove Corollary 13, consider $y = Mx$ and $y = Fz$. From Section 2.2, we may assume M is full rank, and since it has more rows than columns by hypothesis, as pointed out after (1.1), there exists a square invertible matrix, A, such that $x = Az$. In addition, since x and z have a regular covariance, it also holds that $MA = F$. Then, if Ψ is discriminating, it implies by Definition 6 that $A = \Lambda P$, where Λ is a scaling diagonal matrix and P is a permutation. Premultiplying by M this last equality eventually gives $F = M\Lambda P$. □

7.A.10. Sketch of the proof for (3.4)

Start with the expansion

$$(1+v)\log(1+v) = v + v^2/2 - v^3/6 + v^4/12 + o(v^4). \quad \text{(A.10)}$$

Let the relation $p_z(x) = \phi(x)(1 + v(x))$. Then $v(x)$ is obviously defined through relation (3.2), and the negentropy (2.13) can be written as

$$J(p_z) = \int \phi(x)(1+v(x))\log(1+v(x))\,dx. \quad \text{(A.11)}$$

Inserting (A.10) into (A.11), and then replacing $v(x)$ by its value given by (3.2), leads to a long expression. In this expression, a number of simplifications can be performed by resorting to the following properties of Hermite polynomials.

- orthogonality with respect to the Gaussian kernel:

$$\int \phi(u)h_p(u)h_q(u)\,du = p!\,\delta_{pq}; \quad \text{(A.12)}$$

- other less known properties that can be proved by successive integrations by parts:

$$\int \phi(u)h_3^2(u)h_4(u)\,du = 3!^3, \quad \text{(A.13)}$$

$$\int \phi(u)h_3^2(u)h_6(u)\,du = 6!, \quad \text{(A.14)}$$

$$\int \phi(u)h_3^4(u)\,du = 93\cdot 3!^2. \quad \text{(A.15)}$$

Then retaining only the terms at least of order $O(m^{-2})$, according to (3.1), yields finally (3.4). □

7.A.11. Proof of (3.5)

If $z = Ay$, then

$$S(p_y) = -\int p_z(Au)\log\{|A|p_z(Au)\}|A|\,du, \quad \text{(A.16)}$$

which yields $S(p_y) = S(p_z) - |A|\log|A|$. When A is orthogonal, $|A| = 1$ and $S(p_y) = S(p_z)$ as well as $S(\phi_y) = S(\phi_z)$. □

7.A.12. Proof of Lemma 15

Lemma 15 and Theorem 16 also hold for unitary matrices. Since the proof has the same complexity, we shall derive it in this case. Denote by $\bar{Q}$ the matrix obtained by replacing all its entries by their modulus. Consider a unitary matrix Q; then the matrix ${}^2\bar{Q}$ is bistochastic (i.e. the sum of its entries in any row or column is equal to 1). Yet from the Birkhoff theorem, the set of bistochastic matrices is a convex polyhedron whose vertices are permutations. Thus, ${}^2\bar{Q}$ can be decomposed as

$${}^2\bar{Q} = \sum_s \alpha_s P_s, \quad \alpha_s \geq 0, \ \sum_s \alpha_s = 1, \quad \text{(A.17)}$$

where P_s are permutation matrices. Denote by $\bar{u}$ the vector of components $|u_i|$. Then we have the inequality

$$\|{}^2Qu\| \leq \|{}^2\bar{Q}\bar{u}\| \leq \sum_s \alpha_s \|P_s\bar{u}\| = \|u\|. \quad \square \quad \text{(A.18)}$$

7.A.13. Proof of Theorem 16

Since Q is orthogonal, we have $|Q_{ij}| \leq 1$ and, consequently, $|{}^rQ_{ij}| \leq |{}^2Q_{ij}|$ as soon as $r \geq 2$, which can also be written as ${}^r\bar{Q}_{ij} \leq {}^2\bar{Q}_{ij}$. Applying the triangular inequality gives

$$\sum_{i,j,k} Q_{ki}^r Q_{kj}^r u_i u_j \leq \sum_{i,j,k} {}^r\bar{Q}_{ki}\,{}^r\bar{Q}_{kj}\bar{u}_i\bar{u}_j \leq \sum_{i,j,k} {}^2\bar{Q}_{ki}\,{}^2\bar{Q}_{kj}\bar{u}_i\bar{u}_j. \quad \text{(A.19)}$$

Then using Lemma 15 we get

$$\|{}^rQu\|^2 \leq \|{}^2\bar{Q}\bar{u}\|^2 \leq \|\bar{u}\|^2 = \|u\|^2. \quad \text{(A.20)}$$

Now consider a standardized random vector $\tilde{y}$ having zero cross-cumulants of order r and denote by u_i its marginal cumulants of order r. Because of the multilinearity of cumulants (e.g. (3.7) or (3.8)), the very left-hand side of (A.20) is nothing else but $\Psi(Q\tilde{y})$. So we have just proved with (A.20) that $\Psi(Q\tilde{y}) \leqslant \Psi(\tilde{y})$ for any orthogonal matrix Q. This shows that Theorem 16 is indeed a contrast, as soon as $r \geqslant 2$.

Note that, for $r = 2$, the equality $\|^2\bar{Q}\bar{u}\| = \|u\|$ always holds, so that the contrast is actually degenerated. It remains to prove that Theorem 16 is discriminating for any $r > 2$.

Assume equality in (A.20). Then, in particular,

$$\|^2\bar{Q}\bar{u}\|^2 - \|^r\bar{Q}\bar{u}\|^2 = 0 \tag{A.21}$$

for some vector u having at most one null component. But because of (A.19), this implies that all the differences involved in (A.21) individually vanish:

$$(^2\bar{Q}\bar{u})_i^2 - (^r\bar{Q}\bar{u})_i^2 = 0, \quad \forall i.$$

Again, because $\bar{u}_j, {}^r\bar{Q}_{ij}$ and ${}^2\bar{Q}_{ij}$ are positive, this yields

$$(^2\bar{Q}\bar{u})_i - (^r\bar{Q}\bar{u})_i = 0, \quad \forall i, \tag{A.22}$$

and next

$$(^2\bar{Q}_{ij} - {}^r\bar{Q}_{ij})\bar{u}_j = 0, \quad \forall i, j. \tag{A.23}$$

Consequently, for all values of i, and for at least $p - 1$ values of j, we have $\bar{Q}_{ij}^2 = \bar{Q}_{ij}^r$. Since $0 \leqslant \bar{Q}_{ij} \leqslant 1$, ${}^2\bar{Q}_{ij}$ is necessarily either zero or one for those pairs (i, j). Now, a $p \times p$ doubly stochastic matrix ${}^2\bar{Q}$ containing only zeros or ones in a $p \times (p-1)$ submatrix is a permutation. The matrix Q is thus equal to a permutation up to multiplication of numbers of unit modulus. This may be written as $Q = DP$, where D need only be diagonal with unit modulus entries. This proves the second part of the theorem.

This result also holds in the complex case, as demonstrated in [15], provided the following definition is used in Theorem 16: $\psi(Q) = \sum|K_{ii\ldots i}|^2$. □

7.A.14. Proof of (3.10)

Since only standardized cumulants are involved, we only need consider orthogonal transforms. Again from the multilinearity of cumulants, we have

$$\Omega_r = \sum_{i_1 \ldots i_r} \sum_{p_1 \ldots p_r} \sum_{q_1 \ldots q_r} Q_{i_1 p_1} \cdots Q_{i_r p_r} \, Q_{i_1 q_1} \cdots Q_{i_r q_r} \Gamma_{p_1 \ldots p_r} \Gamma_{q_1 \ldots q_r}, \tag{A.24}$$

Now put the first sum inside the two others, and remember that because Q is orthogonal we have

$$\sum_i Q_{ip} Q_{iq} = \delta_{pq}.$$

Then (A.24) simplifies into

$$\Omega_r = \sum_{p_1 \ldots p_r} \sum_{q_1 \ldots q_r} \delta_{p_1 q_1} \cdots \delta_{p_r q_r} \Gamma_{p_1 \ldots p_r} \Gamma_{q_1 \ldots q_r}, \tag{A.25}$$

which is eventually nothing else but the sum

$$\Omega_r = \sum_{p_1 \ldots p_r} \Gamma^2_{p_1 \ldots p_r},$$

which does not depend on matrix Q. These results also hold in the complex case, as shown in [15]. □

7.A.15. Proof of (3.11)

$\Omega_{3,4}$ can be shown to be invariant in exactly the same way as the proof of (3.10) was derived. In fact, replace cumulants K as functions of cumulants Γ and matrix Q, using the multilinearity (3.7) and (3.8). Note that if a row index of Q, say i, is present in a monomial of (3.11) then it appears twice. The sum over i can be pulled and calculated first. Then the two terms, say Q_{ia} and Q_{ib}, disappear from (3.11) since, by orthogonality of Q,

$$\sum_i Q_{ia} Q_{ib} = \delta_{ab}. \tag{A.26}$$

Then repeat the procedure as long as entries of Q are left in the expression of $\Omega_{3,4}$. This finally gives the expression

$$\Omega_{3,4} = 4 \sum_{abc} \Gamma^2_{abc} + \sum_{abcd} \Gamma^2_{abcd} +$$

$$3\sum_{abcdef}\Gamma_{abc}\Gamma_{abd}\Gamma_{cef}\Gamma_{efd} + 4\Gamma_{abc}\Gamma_{ade}\Gamma_{bdf}\Gamma_{cef} - 6\sum_{abcde}\Gamma_{abc}\Gamma_{ade}\Gamma_{bcde}, \tag{A.27}$$

which proves the invariance of $\Omega_{3,4}$. Note that this would also be true if Q were complex unitary, if squares were replaced by squared moduli. □

7.A.16. Proof of Theorem 17

It is sufficient to look at the derivatives of a monomial. So define

$$\psi(Q) = \sum_i \prod_{\alpha\in S} K_{\alpha^* i}, \tag{A.28}$$

where the notation $\alpha^* i$ stands for a repetition of index i α times and S is a finite set of integers. The differentiation yields

$$\mathrm{d}\psi(Q) = \sum_i \sum_{\alpha\in S} \mathrm{d}K_{\alpha^* i} \prod_{\beta\neq\alpha} K_{\beta^* i}. \tag{A.29}$$

Yet, we have two relations below:

$$\mathrm{d}K_{\alpha^* i} = \alpha \sum_j \mathrm{d}Q_{ij} K_{j,(\alpha-1)^* i}, \tag{A.30}$$

$$\mathrm{d}Q = \mathrm{d}A\, Q, \tag{A.31}$$

for some skew-symmetric matrix $\mathrm{d}A$. Stationary points of $\psi(Q)$ just cancel the derivative for all perturbations $\mathrm{d}Q$, and thus from (A.31) for all skew-symmetric matrices $\mathrm{d}A$. Write any skew-symmetric matrix as a linear combination of matrices $A(p,q)$ defined by

$$A(p,q)_{ij} = \delta_{pi}\delta_{qj} - \delta_{pj}\delta_{qi}.$$

Consequently, the cancellation of (A.29) is equivalent to

$$\sum_{\alpha\in S} \alpha\Big(K_{q,(\alpha-1)^* p}\prod_{\beta\neq\alpha} K_{\beta^* p} - K_{p,(\alpha-1)^* q}\prod_{\beta\neq\alpha} K_{\beta^* q}\Big) = 0, \quad \forall p, q. \tag{A.32}$$

Reasoning in the same way for $\mathrm{d}^2\psi(Q)$ shows that, at any stationary point, $\mathrm{d}^2\psi$ contains only terms of the form

$$K_{\alpha^* j}, \quad K_{i,(\alpha-1)^* j} \quad \text{and} \quad K_{2^* i,(\alpha-2)^* j}, \tag{A.33}$$

in which still only two distinct indices enter. In fact, let us give just the expression of $\mathrm{d}^2\psi$ at stationary points:

$$\begin{aligned}\sum_{\alpha\in S}\alpha\Big\{&(\alpha-2)K_{2q,(\alpha-2)^* p}\prod_{\beta\neq\alpha}K_{\beta^* p} + K_{q,(\alpha-1)^* p}\prod_{\beta\neq\alpha}\beta K_{q,(\beta-1)^* p}\prod_{\gamma\neq\beta}K_{\gamma^* p}\\ &-(\alpha-1)\prod_{\alpha}K_{\alpha^* p} + (\alpha-2)K_{2p,(\alpha-2)^* q}\prod_{\beta\neq\alpha}K_{\beta^* q} + K_{p,(\alpha-1)^* q}\prod_{\beta\neq\alpha}\beta K_{p,(\beta-1)^* p}\prod_{\gamma\neq\beta}K_{\gamma^* q}\\ &-(\alpha-1)\prod_{\alpha}K_{\alpha^* q}\Big\}.\end{aligned} \tag{A.34}$$

As a conclusion, $\mathrm{d}\psi$ as well as $\mathrm{d}^2\psi$ involve only pairwise cumulants. This proof would actually extend to $\mathrm{d}^n\psi$. Let us end with an example. Suppose $\psi(Q) = \sum_i K_{iii}^2 K_{iiii}$; then $S = \{4,3,3\}$ and (A.32) gives

$$\mathrm{d}\psi = 0 \Leftrightarrow 4(K_{pppq}K_{ppp}^2 - K_{pqqq}K_{qqq}^2) + 6(K_{ppq}K_{ppp}K_{pppp} - K_{pqq}K_{qqq}K_{qqqq}) = 0$$

$\forall p, q$. □

7.A.17. Proof of the contrast expression (4.2)

The proof is a little tedious but raises no difficulty. Assume that the pair of components to be processed is (1, 2) without restricting generality, and start with the definition of the contrast:

$$\Psi(p_z) = K_{1111}^2 + K_{2222}^2. \tag{A.35}$$

Next substitute back in (A.35) expressions of K_{iiii} as functions of θ, taking (3.8) and (4.1) into account:

$$\begin{aligned}(1+\theta^2)^2 K_{1111} &= \Gamma_{2222}\theta^4 + 4\Gamma_{1222}\theta^3 + 6\Gamma_{1122}\theta^2 + 4\Gamma_{1112}\theta + \Gamma_{1111},\\ (1+\theta^2)^2 K_{2222} &= \Gamma_{1111}\theta^4 - 4\Gamma_{1112}\theta^3 + 6\Gamma_{1122}\theta^2 - 4\Gamma_{1222}\theta + \Gamma_{2222},\end{aligned} \tag{A.36}$$

and group the terms such that the expression depends only on $\xi = \theta - 1/\theta$. For the sake of simplicity, denote

$$a_0 = \Gamma_{1111}, \quad a_1 = 4\Gamma_{1112}, \quad a_2 = 6\Gamma_{1122}, \quad a_3 = 4\Gamma_{1222}, \quad a_4 = \Gamma_{2222}. \tag{A.37}$$

Then the contrast (A.35) is of the form (4.2) if we denote the coefficients as follows:

$$
\begin{aligned}
b_4 &= a_0^2 + a_4^2,\\
b_3 &= 2(a_3a_4 - a_0a_1),\\
b_2 &= 4a_0^2 + 4a_4^2 + a_1^2 + a_3^2 + 2a_0a_2 + 2a_2a_4,\\
b_1 &= 2(-3a_0a_1 + 3a_3a_4 + a_1a_4 - a_0a_3 \\
&\quad + a_2a_3 - a_1a_2)\\
b_0 &= 2(a_0^2 + a_1^2 + a_2^2 + a_3^2 + a_4^2 + 2a_0a_2 \\
&\quad + 2a_0a_4 + 2a_1a_3 + 2a_2a_4). \qquad \square
\end{aligned}
\tag{A.38}
$$

7.A.18. Exact form of Eq. (4.4) defining stationary points

Taking the derivative of $\psi(\xi)$ given in (4.2) with respect to ξ yields directly (4.4) if we denote

$$
\begin{aligned}
c_4 &= \Gamma_{1111}\Gamma_{1112} - \Gamma_{2222}\Gamma_{1222},\\
c_3 &= v - 4(\Gamma_{1112}^2 + \Gamma_{1222}^2)\\
&\quad - 3\Gamma_{1122}(\Gamma_{1111} + \Gamma_{2222}),\\
c_2 &= 3\upsilon,\\
c_1 &= 3v - 2\Gamma_{1111}\Gamma_{2222} - 32\Gamma_{1112}\Gamma_{1222}\\
&\quad - 36\Gamma_{1122}^2,\\
c_0 &= -4(\upsilon + 4c_4),
\end{aligned}
\tag{A.39}
$$

with

$$
\begin{aligned}
\upsilon &= (\Gamma_{1111} + \Gamma_{2222} - 6\Gamma_{1122})(\Gamma_{1222} - \Gamma_{1112}),\\
v &= \Gamma_{1111}^2 + \Gamma_{2222}^2.
\end{aligned}
$$

7.A.19. Extension to the complex case

In the complex case, the characteristic function can be defined in the standard way, i.e. $\varphi_y(\boldsymbol{u}) = E\{\exp[i\,\mathrm{Re}(\boldsymbol{y}^*\boldsymbol{u})]\}$, but cumulants can take various forms. Here we use $\Gamma_{ijkl} = \mathrm{cum}\{\tilde{y}_i, \tilde{y}_j^*, \tilde{y}_k^*, \tilde{y}_l\}$ as a definition of cumulants. We consider plane rotations with real cosine and complex sine:

$$
Q = c\begin{bmatrix} 1 & \theta \\ -\theta^* & 1 \end{bmatrix}, \quad c \in \mathbb{R}^+, \quad c = 1/\sqrt{1 + |\theta|^2},
$$

$$
\theta = s/c, \quad c^2 + |s|^2 = 1. \tag{A.40}
$$

Output cumulants given in the real case in (A.36) now take the following form [15]:

$$
\begin{aligned}
K_{1111}/c^4 &= \Gamma_{2222}\theta^2\theta^{*2} + 2\theta\theta^*\{\Gamma_{2122}\theta + \Gamma_{1222}\theta^*\}\\
&\quad + 4\Gamma_{1122}\theta\theta^* + \{\Gamma_{2112}\theta^2 + \Gamma_{1221}\theta^{*2}\}\\
&\quad + 2\{\Gamma_{1112}\theta + \Gamma_{1121}\theta^*\} + \Gamma_{1111},
\end{aligned}
\tag{A.41}
$$

$$
\begin{aligned}
K_{2222}/c^4 &= \Gamma_{1111}\theta^2\theta^{*2} - 2\theta\theta^*\{\Gamma_{1112}\theta + \Gamma_{1121}\theta^*\}\\
&\quad + 4\Gamma_{1122}\theta\theta^* + \{\Gamma_{2112}\theta^2 + \Gamma_{1221}\theta^{*2}\}\\
&\quad - 2\{\Gamma_{2122}\theta + \Gamma_{1222}\theta^*\} + \Gamma_{2222}.
\end{aligned}
$$

Then the contrast function, which is real by construction, can be calculated as a function of the variable $\xi = \theta - 1/\theta^*$, since ψ is a rational function satisfying $\psi(1/\theta^*) = \psi(\theta)$. It can be shown that the precise form of $\psi(\xi)$ is

$$
\psi(\xi) = \sum_{\substack{i=-2 \\ 0 \le i+j \le 4}}^{2} \sum_{j=-2}^{2} b_{ij}\xi^i\xi^{*j}/[\xi\xi^* + 4]^2, \tag{A.42}
$$

where the coefficients b_{ij} are defined as follows. Denote for the sake of simplicity

$$
\begin{aligned}
&a_0 = \Gamma_{1111}, \quad a_1 = 4\Gamma_{1112}, \quad a_2 = \Gamma_{1122},\\
&\tilde{a}_2 = 2\Gamma_{2112}, \quad a_3 = 4\Gamma_{1222}, \quad a_4 = \Gamma_{2222}.
\end{aligned}
\tag{A.43}
$$

Then

$$
\begin{aligned}
b_{22} &= a_0^2 + a_4^2,\\
b_{21} &= a_4a_3^* - a_0a_1, \qquad b_{12} = b_{21}^*,\\
b_{11} &= 4(a_0^2 + a_4^2) + (a_1a_1^* + a_3a_3^*)/2\\
&\quad + 2a_2(a_0 + a_4),\\
b_{20} &= (a_1^2 + a_3^{*2})/4 + \tilde{a}_2(a_0 + a_4), \quad b_{02} = b_{20}^*,\\
b_{10} &= a_2(a_3^* - a_1) + \tilde{a}_2(a_3 - a_1^*)/2\\
&\quad + 3(a_4a_3^* - a_0a_1) + (a_4a_1 - a_0a_3^*),\\
b_{01} &= b_{10}^*,
\end{aligned}
\tag{A.44}
$$

$$b_{2-1} = \tilde{a}_2(a_3^* - a_1)/2, \qquad b_{-12} = b_{2-1}^*,$$

$$b_{1-1} = 2a_2\tilde{a_2} + (a_1^2 + a_3^2)/2 + a_1a_3^* + 2\tilde{a}_2(a_0 + a_4), \qquad b_{-11} = b_{1-1}^*,$$

$$b_{2-2} = \tilde{a}_2^2/2, \qquad b_{-22} = b_{2-2}^*,$$

$$b_0 = 2a_0^2 + a_1a_1^* + \tilde{a}_2\tilde{a}_2^* + 2a_2^2 + a_3a_3^* + 2a_4^2 + 4a_0a_4 + a_1a_3 + a_1^*a_3^* + 4a_2(a_0 + a_4).$$

The other coefficients are null. It may be checked that this expression of the contrast coincides with (A.38) when all the quantities involved are real.

Next, stationary points of ψ can be calculated in various ways. One possibility is to resort to the equation

$$K_{1111}K_{1112} - K_{1222}K_{2222} = 0, \tag{A.45}$$

which is satisfied at any stationary point of contrast (A.35) (relation (A.32) indeed extends to the complex case). The other possibility is to differentiate (A.42) with respect to the modulus ρ and argument φ of ξ, respectively, and set the derivatives to zero. We obtain a system of two polynomials of two variables, whose degree in ρ is equal to 4 and 3, respectively. The exact form of the system obtained is not given here for reasons of space. It can be solved by first performing an elimination on ρ using successive greatest common divisors of coefficients (which are polynomials in $e^{i\varphi}$). Next, rooting the polynomial in $e^{i\varphi}$ alone gives admissible values of φ. Then substitution in the original system allows one to obtain the corresponding admissible values of ρ. As in the real case, the pairs (ρ, φ) corresponding to maxima are selected by replacing ξ by $\rho e^{i\varphi}$ in the expression of the contrast.

7.A.20. Proof of (5.2) and (5.3)

Assume $\hat{A} = A\Lambda P$. Then by definition the gap is a function of the matrix $D' = P^*D$ if we denote $D = \underline{A}^{-1}\underline{\hat{A}}$ (in fact, $\underline{A\Lambda P} = \underline{A\Lambda}P = \underline{A}P$). However, a glance at (5.1) suffices to see that the order in which the rows of D are set does not change the expression of the gap $\varepsilon(A, \hat{A})$.

Let us prove the converse implication. If $\varepsilon(A, \hat{A}) = 0$, then each term in (5.1) must vanish; this yields

$$\text{(a)} \sum_j |D_{ij}| = 1, \quad \text{(b)} \sum_i |D_{ij}| = 1,$$

$$\text{(c)} \sum_j |D_{ij}|^2 = 1, \quad \text{(d)} \sum_i |D_{ij}|^2 = 1. \tag{A.46}$$

The first two relations show that matrix $\bar{D}$ is bistochastic. Yet, we have shown in Appendix A.13 that for any bistochastic matrix $\bar{D}$, if there exists a vector $\bar{u}$ with real positive entries and at most one null component such that

$$\|\bar{D}\bar{u}\| = \|^2\bar{D}\bar{u}\|, \tag{A.47}$$

then $\bar{D}$ is a permutation. Here the result we have to prove is weaker. In fact, choose as vector $\bar{u}$ the vector formed of ones. From (A.46) we have

$$\bar{D}\bar{u} = {}^2\bar{D}\bar{u} = 1, \tag{A.48}$$

which of course implies (A.47). Thus, $\bar{D}$ is a permutation, and $\hat{A}$ is of the expected form. □

7.A.21. Description of the 'validation' algorithm

This algorithm aims to provide ultimate performances when statistics are perfectly known. It is supposed that M is known as well as the vector of source and noise cumulants, denoted by κ_{pppp} in this appendix, $1 \leqslant p \leqslant 2N$. The algorithm differs from Algorithm 18 only in that moments are deduced from the current transformation and the source cumulants.

ALGORITHM 21 (Validation).

(1) Compute the SVD of the matrix $AA = [(1 - \mu)M \ \mu\beta I]$ as $AA = V\Sigma U^*$, where V is $N \times \rho$, Σ is $\rho \times \rho$ and U^* is $\rho \times 2N$, and ρ denotes the rank of AA. Set $L = V\Sigma$.

(2) Initialize $F = L$.

(3) Begin a loop on the sweeps: $k = 1, 2, \ldots, k_{max}$; $k_{max} \leqslant 1 + \sqrt{\rho}$.

(4) Sweep the $\rho(\rho - 1)/2$ pairs (i, j), according to a fixed ordering. For each pair:

(a) Use the multilinearity relation (3.8) in order to compute the required cumulants $\Gamma_{\alpha(i,j)}$ as functions of cumulants κ_{pppp} and matrices AA and F.
(b) Find the angle α maximizing $\psi(Q^{(i,j)})$, where $Q^{(i,j)}$ is the plane rotation of angle α, $\alpha \in]-\pi/4, \pi/4]$, in the plane defined by components $\{i, j\}$. Use the expressions given in Appendices A.17 and A.18 for this purpose.
(c) Accumulate $F := FQ^{(i,j)*}$.

(5) End of the loop on k if $k = k_{\max}$, or if all estimated angles are small (compared to machine precision).
(6) Compute the norm of the columns of F: $\Delta_{ii} = \| F_{ii} \|$.
(7) Sort the entries of Δ in decreasing order:

$$\Delta := P\Delta P^* \text{ and } F := FP^*.$$

(8) Normalize F by the transform $F := F\Delta^{-1}$.
(9) Fix the phase (sign) of each column of F according to Definition 4(e). This yields $F := FD$.

Source MATLAB codes can be obtained from the author upon request. □

8. Acknowledgments

The author wishes to thank Albert Groenenboom, Serge Prosperi and Peter Bunn for their proofreading of the paper, which unquestionably improved its clarity. References [44], indicated by Ananthram Swami, and [6], indicated by Albert Benveniste some years ago, were also helpful. Lastly, the author is indebted to the reviewers, who detected a couple of mistakes in the original submission, and permitted a definite improvement in the paper.

9. References

[1] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1965, 1972.
[2] Y. Bar-Ness, "Bootstrapping adaptive interference cancelers: Some practical limitations", *The Globecom Conf.*, Miami, November 1982, Paper F3.7, pp. 1251–1255.
[3] S. Bellini and R. Rocca, "Asymptotically efficient blind deconvolution", *Signal Processing*, Vol. 20, No. 3, July 1990, pp. 193–209.
[4] M. Basseville, "Distance measures for signal processing and pattern recognition", *Signal Processing*, Vol. 18, No. 4, December 1989, pp. 349–369.
[5] A. Benveniste, M. Goursat and G. Ruget, "Robust identification of a nonminimum phase system", *IEEE Trans. Automat. Control*, Vol. 25, No. 3, 1980, pp. 385–399.
[6] R.E. Blahut, *Principles and Practice of Information Theory*, Addison-Wesley, Reading, MA, 1987.
[7] D.R. Brillinger, *Time Series Data Analysis and Theory*, Holden-Day, San Francisco, CA, 1981.
[8] J.F.Cardoso, "Sources separation using higher order moments", *Proc. Internat. Conf. Acoust. Speech Signal Process.*, Glasgow, 1989, pp. 2109–2112.
[9] J.F. Cardoso, "Super-symmetric decomposition of the fourth-order cumulant tensor, blind identification of more sources than sensors", *Proc. Internat. Conf. Acoust. Speech Signal Process. 91*, 14–17 May 1991.
[10] J.F. Cardoso, "Iterative techniques for blind sources separation using only fourth order cumulants", *Conf. EUSIPCO*, 1992, pp. 739–742.
[11] T.F. Chan, "An improved algorithm for computing the singular value decomposition", *ACM Trans Math. Software*, Vol. 8, No. March 1982, pp. 72–83.
[12] P. Comon, "Separation of sources using high-order cumulants", *SPIE Conf. on Advanced Algorithms and Architectures for Signal Processing, Vol. Real-Time Signal Processing XII*, San Diego, 8–10 August 1989, pp. 170–181.
[13] P. Comon, "Separation of stochastic processes", *Proc. Workshop Higher-Order Spectral Analysis*, Vail, Colorado, June 1989, pp. 174–179.
[14] P. Comon, "Process and device for real-time signals separation", Patent registered for Thomson-Sintra, No. 9000436, December 1989.
[15] P. Comon, "Analyse en Composantes Indépendantes et Identification Aveugle", *Traitement du Signal*, Vol. 7, No. 5, December 1990, pp. 435–450.
[16] P. Comon, "Independent component analysis", *Internat. Signal Processing Workshop on High-Order Statistics*, Chamrousse, France, 10–12 July 1991, pp. 111–120; republished in J.L. Lacoume, ed., *Higher-Order Statistics*, Elsevier, Amsterdam 1992, pp. 29–38.
[17] P. Comon, "Kernel classification: The case of short samples and large dimensions", Thomson ASM Report, 93/S/EGS/NC/228.
[18] H. Cramér, *Random Variables and Probability Distributions*, Cambridge Univ. Press, Cambridge, 1937.
[19] G. Darmois, "Analyse Générale des Liaisons Stochastiques", *Rev. Inst. Internat. Stat.*, Vol. 21, 1953, pp. 2–8.
[20] G. Desodt and D. Muller, "Complex independent component analysis applied to the separation of radar signals", in: Torres, Masgrau and Lagunas, eds., *Proc. EUSIPCO Conf.*, Barcelona, Elsevier, Amsterdam, 1990, pp. 665–668.

[21] D.L. Donoho, "On minimum entropy deconvolution", *Proc. 2nd Appl. Time Series Symp.*, Tulsa, 1980; reprinted in *Applied Time Series Analysis II*, Academic Press, New York, 1981, pp. 565-608.

[22] D. Dugué, "Analycité et convexité des fonctions caractéristiques", *Annales de l'Institut Henri Poincaré*, Vol. XII, 1951, pp. 45-56.

[23] P. Duvaut, "Principes de méthodes de séparation fondées sur les moments d'ordre supérieur", *Traitement du Signal*, Vol. 7, No. 5, December 1990, pp. 407-418.

[24] L. Fety, Methodes de traitement d'antenne adaptées aux radiocommunications, Doctorate Thesis, ENST, 1988.

[25] P. Flandrin and O. Michel, "Chaotic signal analysis and higher order statistics", *Conf. EUSIPCO*, Brussels, 24-27 August 1992, pp. 179-182.

[26] P. Forster, "Bearing estimation in the bispectrum domain", *IEEE Trans. Signal Processing*, Vol. 39, No. 9, September 1991, pp. 1994-2006.

[27] J.C. Fort, "Stability of the JH sources separation algorithm", *Traitement du Signal*, Vol. 8, No. 1, 1991, pp. 35-42.

[28] B. Eriedlander and B. Porat, "Asymptotically optimal estimation of MA and ARMA parameters of non-Gaussian processes from high-order moments", *IEEE Trans Automat. Control*. Vol. 35, January 1990, pp. 27-35.

[29] M. Gaeta, Les Statistiques d'Ordre Supérieur Appliquées à la Séparation de Sources, Ph.D. Thesis, Université de Grenoble, July 1991.

[30] M. Gaeta and J.L. Lacoume, "Source separation without a priori knowl'edge: The maximum likelihood solution", in: Torres, Masgrau and Lagunas, eds., *Proc. EUSIPCO Conf.*, Barcelona, Elsevier, Amsterdam, 1990, pp. 621-624.

[31] E. Gassiat, Déconvolution Aveugle, Ph.D. Thesis, Université de Paris Sud, January 1988.

[32] G. Giannakis and S. Shamsunder, "Modeling of non-Gaussian array data using cumulants: DOA estimation with less sensors than sources", *Proc. 25th Ann. Conf. on Information Sciences and Systems*, Baltimore, 20-22 March 1991, pp. 600-606.

[33] G. Giannakis and A. Swami, "On estimating noncausal nonminimum phase ARMA models of non-Gaussian processes", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 38. No. 3, March 1990, pp. 478-495.

[34] G. Giannakis and M.K. Tsatsanis, "A unifying maximum-likelihood view of cumulant and polyspectral measures for non-Gaussian signal classification and estimation", *IEEE Trans. Inform. Theory*, Vol. 38, No. 2, March 1992, pp. 386-406.

[35] G. Giannakis, Y. Inouye and J.M. Mendel, "Cumulant-based identification of multichannel moving average models", *IEEE Automat. Control*, Vol. 34, July 1989, pp. 783-787.

[36] Y. Inouye and T. Matsui, "Cumulant based parameter estimation of linear systems", *Proc. Workshop Higher-Order Spectral Analysis*, Vail, Colorado, June 1989, pp. 180-185.

[37] C. Jutten and J. Herault, "Blind separation of sources, Part I: An adaptive al'gorithm based on neuromimatic architecture", *Signal Processing*, Vol. 24, No. 1, July 1991, pp. 1-10.

[38] A.M. Kagan, Y.V. Linnik and C.R. Rao, *Characterization Problems in Mathematical Statistics*, Wiley, New York, 1973.

[39] M. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Vol. 1, New York, 1977.

[40] S. Kotz and N.L. Johnson, *Encyclopedia of Statistical Sciences*, Wiley, New York, 1982.

[41] J.L. Lacoume, "Tensor-based approach of random variables", Talk given at ENST Paris, 7 February 1991.

[42] J.L. Lacoume and P. Ruiz, "Extraction of independent components from correlated inputs, A solution based on cumulants", *Proc Workshop Higher-Order Spectral Analysis*, Vail, Colorado, June 1989, pp. 146-151.

[43] P. McCullagh, "Tensor notation and cumulants", *Biometrika*, Vo. 71, 1984, pp. 461-476.

[44] P. McCullagh, *Tensor Methods in Statistics*, Chapman and Hall, London 1987.

[45] C.L. Nikias, "ARMA bispectrum approach to nonminimum phase system identification", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 36, April 1988, pp. 513-524.

[46] C.L. Nikias and M.R. Raghuveer, "Bispectrum estimation: A digital signal processing framework", *Proc. IEEE*, Vol. 75, No. 7, July 1987, pp. 869-891.

[47] B. Porat and B. Friedlander, "Direction finding algorithms based on higher-order statistics", *IEEE Trans. Signal Processing*, Vol. 39, No. 9, September 1991, pp. 2016-2024.

[48] J.G. Proakis and C.L. Nikias, "Blind equalization", in: *Adaptive Signal Processing, Proc. SPIE*, Vol. 1565, 1991, pp. 76-85.

[49] C.R. Rao, *Linear Statistical Inference and its Applications*, Wiley, New York, 1973.

[50] P.A. Regalia and P. Loubaton, "Rational subspace estimation using adaptive lossless filters", *IEEE Trans. Acoust. Speech Signal Process.*, July 1991, submitted.

[51] G. Scarano, "Cumulant series exapansion of hybrid nonlinear moments of complex random variables", *IEEE Trans. Signal Processing*, Vol. 39, No. 4, April 1991, pp. 1001-1003.

[52] O. Shalvi and E. Weinstein, "New criteria for blind deconvolution of nonminimum phase systems", *IEEE Trans. Inform. Theory*, Vol. 36, No. 2, March 1990, pp. 312-321.

[53] S. Shamsunder and G.B. Giannakis, "Modeling of non-Gaussian array data using cumulants: DOA estimation of more sourcdes with less sensors", *Signal Processing*, Vol. 30, No. 3, February 1993, pp. 279-297.

[54] J.E. Shore and R.W. Johnson, "Axiomatic derivation of the principle of maximum entropy", *IEEE Trans. Inform Theory*, Vol. 26, January 1980, pp. 26-37.

[55] A. Souloumiac and J.F. Cardoso, "Comparaisons de Méthodes de Séparation de Sources", *XIIIth Coll. GRETSI*, September 1991, pp. 661-664.

[56] A. Swami and J.M. Mendel, "ARMA systems excited by non-Gaussian processes are not always identifiable", *IEEE Trans. Automat. Control*, Vol. 34, No. 5, May 1989, pp. 572-573.

[57] A. Swami and J.M. Mendel, "ARMA parameter estimation using only output cumulants", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 38, July 1990, pp. 1257–1265.

[58] A. Swami, G.B. Giannakis and S. Shamsunder, "A unified approach to modeling multichannel ARMA processes using cumulants", *IEEE Trans. Signal Process.*, submitted.

[59] L. Swindlehurst and T. Kailath, "Detection and estimation using the third moment matrix", *Internat. Conf. Acoust. Speech Signal Process.*, Glasgow, 1989, pp. 2325–2328.

[60] L. Tong, V.C. Soon and R. Liu, "AMUSE: A new blind identification algorithm", *Proc. Internat. Conf. Acoust. Speech Signal Process.*, 1990, pp. 1783–1787.

[61] L. Tong et al., "A necessary and sufficient condition of blind identification", *Internat. Signal Processing Workshop on High-Order Statistics*, Chamrousse, France, 10–12 July 1991, pp. 261–264.

[62] J.K. Tugnait, "Identification of linear stochastic systems via second and fourth order cumulant matching", *IEEE Trans. Inform. Theory*, Vol. 33, May 1987, pp. 393–407.

[63] J.K. Tugnait, "Approaches to FIR system identification with noise data using higher-order statistics", *Proc. Workshop Higher-Order Spectral Analysis*, Vail, June 1989, pp. 13–18.

[64] J.K. Tugnait, "Comments on new criteria for blind deconvolution of nonminimum phase systems", *IEEE Trans. Inform. Theory*, Vol. 38, No. 1, January 1992, 210–213.

[65] D.L. Wallace, "Asymptotic approximations to distributions", *Ann. Math. Statist.*, Vol. 29, 1958, pp. 635–654.