Blind Techniques

Pierre Comon
Laboratory I3S, Cnrs,
University of Nice
www.i3s.unice.fr/∼pcomon

January 17, 2010

Contents

Preamble

These are the notes associated with a course on Blind Identification and Equalization (D16) given for the STIC Master Program "SiCom" (Signal and Digital Communications) at Sophia-Antipolis between 2002 and 2007, belonging to the education program of the University of Nice. Part of this course is also given in 2005 at the Graduate School in Systems and Control, Leuven, Belgium. A shorter version of these notes have been given to the participants in the ICASSP 2005 tutorial #TUT10. A superficial update of these 2005 notes has been performed in January 2010 by adding a few bibliographical pointers and a subject index.

1 Summary

The problem of identifying linear mixtures of independent random variables only from outputs can be traced back to 1953 with the works of Darmois or Skitovich. They pointed out that when data are non Gaussian, a lot more can be said about the mixture.

In practice, Blind Identification of linear mixtures is useful especially in Factor Analysis, in addition to many other application areas (including signal and image processing, digital communications, biomedical, or complexity theory). Harshman and Carroll provided independently numerical algorithms to decompose a data record stored in a 3-way array into elementary "decomposable" arrays, each representing the contribution of a single underlying factor. In [75], the multilinear model had been called "Parafac". Carrol used in [24] the more appropriate terminology of "Canonical Decomposition". Another terminology, introduced by [77] is "Polyadic Decomposition". For these reasons, this tensor decompositioon is often referred to as the "CP" decomposition.

The main difference with the well known Principal Component Analysis is that the mixture is not imposed to be a unitary matrix. This is very relevant because the actual mixture often has no reason to have orthogonal columns. The *Alternating Least Squares* (ALS) algorithm, widely used since that time, theoretically does not converge for topological reasons [46], but yields very usable results after a finite number of iterations under mild conditions.

Independently, the problem of *Blind Source Separation* (BSS) arose around 1985 and was solved -explicitly or implicitly- with the help of High-Order Statistics (HOS), which are actually tensor objects (see [44, Ch.1] for an historical survey). It gave rapidly birth to the more general problem of *Independent Component Analysis* (ICA) in 1991. ICA is a tool that can be used to extract factors when the physical diversity does not allow to store efficiently the data in tensor format, in other words when the multilinear model cannot be used.

This tutorial provides a very accessible background on Statistical Independence, High-Order Statistics, and Tensors. Simple examples are given throughout the talk in order to illustrate various concepts and properties. It emphasizes both the usefulness and limitations of multilinear models and ICA algorithms. Mathematically advanced topics are only tackled, but striking tensor properties that are not satisfied by matrices are still touched upon. Some reported results show how strange and attractive this research area can be.

2 Introduction

2.1 Origins of the subject

Blind Techniques (BT) are born around the seventies, when the first adaptive blind equalizers have been devised in Digital Communications [124] [67] [12]. Up to recently, channel equalization was performed

with the help of *pilot sequences*, which are known to the receiver, and regularly transmitted. The use of pilot sequences have certain drawbacks.

The first one is the most obvious, and is simply the limitation of *throughput* (transmission rate), due to the assignment of part of the symbols to imposed values. Another limitation lies in the fact that it is difficult to take fully advantage of pilot sequences when they are shorter than channel impulse responses, or when channels are fast varying with time. On the other hand, blind equalizers do not use the possible prior knowledge of some symbols in the transmitted bursts, and only exploit the fact that underlying data have a discrete distribution, and hence non Gaussian.

Without blind techniques, pilot sequences are mandatory when "the eye is not open", that is, when the Signal to Noise Ratio (SNR) is too weak or when the fading is too strong in order to be able to separate the received symbol from the other interferences present in the observation. In other works, contrary to Decision-Directed Equalizers, *Blind Equalizers* do not need the eye to be open. The terminology is thus meaningful with this respect.

After a closer look at the literature, it seems that Blind Techniques can find their roots in even early results borrowed from statistics, related to linear mixtures of independent random variables [58] [50] [60]. Applications of these basic principles to exploratory analysis [79] or Digital Communications [57] [9] have come later.

Beside motivations driven by Digital Communications, another community has pushed the development of BT. In fact BT, if restricted to static MIMO modeling, are closely related to Factor Analysis as will be demonstrated subsequently. Other roots are thus located in the works of Tucker [143], Harshman [75] and others [24] [89] [77]. Some recent works (e.g. [14] and references therein) have also shown that old results on polynomials in several variables could be useful [123].

It took some years for this subject to raise interest at a much larger scale. The French community is perhaps responsible for the development of Independent Component Analysis (ICA), since they have authored most of the publications appeared between 1980 and 1995; see [84] [31] [21] [33] [23] [35] and references therein. This is of course not true anymore, and a great number of contributions have appeared since these first shy steps.

In order to close this very brief and necessarily incomplete historical survey, let us insist that some books begin to come out, each addressing part of this wide area. One can in particular mention on one hand [132] on Multi-Way Factor Analysis, [90] [76] on Independent Component Analysis (ICA), and on the other hand [113] [117] on High-Order Statistics or [85] on statistical tools.

2.2 Block methods

Blind techniques have opened a new area of research, which has really been investigated only in the nineties, probably because of the increasing computational power of microprocessors. Ten years later, the computational power has so much increased that it can be envisaged to implement *block algorithms*. As opposed to the so-called *adaptive algorithms*, which are actually built on recursions on time, sample by sample (understand: update of a solution at every symbol period), block algorithms process not only a sample (one symbol) but a whole block of samples (symbols) at a time.

There are several advantages in proceeding this way. First, most adaptive algorithms cumulate in a single computation one step of an iterative optimization algorithm, and an implicit update of the optimization criterion. Because of this combination of two different stages, the convergence time of the optimization algorithm and that of the criterion estimation cannot be adjusted separately. This is for instance the case of the so-called *LMS algorithm*, surprisingly still widely used, despite its poor capabilities.

But this is not the worse drawback of adaptive algorithms of this type. In fact, the optimization step is often as simple as just a Newton gradient descent, which can obviously lead the algorithm to get stuck in a local minimum. It is indeed important to keep in mind that many optimization criteria do have many local minima, among which some may yield spurious solutions.

Block algorithms offer, at least theoretically, the possibility to build more fancy optimization procedures and more freedom in adjusting the parameters. In addition, there are the natural way to pose signal processing problems in the frame of *burst-mode* transmissions (Time Division Multiple Access (*TDMA*)).

For instance, in the GSM standard, a mobile receiving at 900 MHz and 190km/h would operate on a channel having a coherence time of less than 2ms, which is of order of 300 symbols only. It is thus very important to fully exploit the information contained in a so short block, and not to waste the data by using a poor numerical algorithm needing hundreds of thousands of symbols to converge in a stationary context.

Block algorithms enjoy other potential advantages, namely a better resistance to loss in synchronization, and the ability to process the data with time reversal. Hence, they offer more capabilities, which may be utilized in advanced algorithms.

2.3 Observation model

2.3.1 Notation

Arrays (as opposed to scalars) will be denoted throughout with bold face letters. More precisely, a $K \times 1$ array will be called a vector, and will be denoted with a lower case bold letter, *e.g.* $\boldsymbol{x}$; matrices and tensors, which are arrays of larger order, will be denoted with upper case bold letters, *e.g.* $\boldsymbol{H}$.

The *order* of an array is the number of its ways, that is, the number of indices that are necessary to describe its entries. A covariance matrix is for instance a two-way array, and its entries are numbered in a natural manner with two indices. It is also the central moment of order 2 [87].

Random processes we shall use are random variables depending on some time index. If the time index is continuous, the dependence will be denoted with parentheses, *e.g.* $\boldsymbol{x}(t)$; for discrete-time processes, the time index will be put within brackets, as $\boldsymbol{x}[n]$, $n \in \mathbb{N}$.

2.3.2 Dynamic modeling

In this course, it will be always assumed that the observation $\boldsymbol{x}[n]$ is the result of a noisy linear filtering of some *source* process $\boldsymbol{s}[n]$. In other words,

$$\boldsymbol{x}[n] = \boldsymbol{H} \star \boldsymbol{s}[n] + \boldsymbol{b}[n] \tag{1}$$

where $\star$ denotes convolution and where, for every n, $\boldsymbol{x}[n]$, $\boldsymbol{s}[n]$ and $\boldsymbol{b}[n]$ are random vectors of dimension K, P, and K, respectively. In addition, the noise term $\boldsymbol{b}[n]$ is assumed to be statistically independent of $\boldsymbol{x}[n']$ for every pair (n, n'). The integers K and P are commonly referred to as the number of sensors, and the number of sources, respectively.

In the model above, $\boldsymbol{H}[n]$ represents the impulse response of the filter (often called the channel in digital communications) linking source and observation; each term $\boldsymbol{H}[n]$ is a $K \times P$ matrix, for every n. For practical reasons, filters will be often assumed to have a Finite Impulse Response, so that $1 \leq n \leq L$.

Note that model (1) is not always appropriate in digital communications. In fact, consider a symbol sequence $s[k]$, $k \in \mathbb{N}$, that we wish to transmit through a channel. For this purpose, we use a continuous-time transmit filter, $g(t)$. The continuous signal transmitted is thus $g \star s(t)$, defines

without ambiguity as:

$$g \star s(t) = \sum_k g(t - kT_s)\, s[k]$$

where T_s denotes the symbol period. After propagation through the channel, and through a receive filter, the received signal takes the similar form:

$$\boldsymbol{y}(t) \stackrel{\text{def}}{=} \boldsymbol{h} \star s(t) = \sum_k \boldsymbol{h}(t - kT_s)\, s[k] \tag{2}$$

where $\boldsymbol{h}(t)$ denotes the *global impulse response*, that is, the composition of the receive filter, the channel, and the transmit filter. If this signal is sampled at a rate $1/T$, the obtained discrete-time process is

$$\boldsymbol{y}[n] = \sum_k \boldsymbol{h}(nT - kT_s)s[k] \tag{3}$$

which cannot, in general, be put in the form of a *discrete convolution*. Additionally, even if process $\boldsymbol{y}[n]$ is the convolution of a linear filter of a stationary process $s[n]$, it is not generally stationary! The only thing we can claim is all the statistical properties of $\boldsymbol{y}(t)$ are periodic of period T_s, which is referred to as the *cyclo-stationarity* property. A sufficient condition for getting a discrete convolution in (3) is that $T = T_s$, which means a perfect symbol rate synchronization. Under this condition, we may use the simplified model below, which may be easily seen to be an instance of model (1):

$$\boldsymbol{y}[n] = \sum_k \boldsymbol{h}[n - k]s[k]$$

where $\boldsymbol{h}[n] \stackrel{\text{def}}{=} \boldsymbol{h}(nT_s)$.

2.3.3 Static modeling

In some applications, this model can be deflated to a static one, namely:

$$\boldsymbol{x} = \boldsymbol{H}\, \boldsymbol{s} + \boldsymbol{b} \tag{4}$$

In such a case, the channel is entirely characterized by a single (generally complex) $K \times P$ matrix. This is in particular the case in digital communications in the presence of flat fading [122]. This model is also the starting point of *Independent Component Analysis* (ICA) [35], now widely used in Factor Analysis, among other numerous applications. The present tutorial is mainly devoted to the static model framework.

2.3.4 Problem formulation

The goal of *Blind Equalization* is to estimate the sequence $\boldsymbol{s}[n]$ from the sole observation of realizations of $\boldsymbol{y}[n]$, which is assumed to satisfy model (1) with a sufficiently high SNR. When the number of sources $P > 1$, this implies to separate the sources.

When the observation model is static, as in (4), the problem is meaningful only when $P > 1$, in which case it reduces to *Blind Source Separation*.

The usage of terminology is such that one may currently understand by "Blind Source Separation" an estimation of individual sources either from the static model (4) or from the dynamic model (1).

Another very different problem is that of *Blind Identification*, which consists of estimating the global channel $\boldsymbol{H}[k]$, but not of extracting the source processes $s_i[n]$. Blind Equalization and Blind Identification are addressed in different ways, and it is important to distinguish between both problems.

At a methodological level, the following questions should be sequentially answered. First, it is needed to define a theoretical optimization criterion, having the relevant properties; it can be based on results borrowed from Information Theory. Then in a second stage, a practical criterion must be derived, possibly from sample moments. Third, numerical algorithms have to be devised in order to perform the practical optimization. Last, performances are accessed through evaluation indices, which can be related directly or indirectly to the optimization criterion. These questions will be the subject of this course.

2.3.5 Inherent Indeterminations

It is clear that, without additional hypotheses, there are infinitely many solutions to the problems of Blind Equalization or Blind Source Separation. In fact, if $\boldsymbol{x} = \boldsymbol{H} \star \boldsymbol{s}$, for some pair $(\boldsymbol{H}, \boldsymbol{s})$, then the equality also holds for any pair $\boldsymbol{H} \star \boldsymbol{T}, \boldsymbol{T}' \star \boldsymbol{s}$, whenever $\boldsymbol{T} \star \boldsymbol{T}' = \boldsymbol{I}$, the identity matrix. Different problems can be addressed, depending on the additional hypotheses assumed (see section 5.1):

- Each source $s_i[n]$ is white in the strong sense, that is, independent and identically distributed (*i.i.d.*).
- Sources $s_i[n]$ and $s_j[m]$ are mutually statistically independent, $\forall i, j,\ n, m$, and at most one source is Gaussian
- Sources are discrete (but might be statistically dependent)
- Each source $\boldsymbol{s}_i[n]$ is colored, and the set of the P spectra forms a family of P linearly independent functions.
- Sources are nonstationary, with linearly independent profiles

The second hypothesis listed above is the most widely utilized to date. Note that in the static case (4), this hypothesis means that the random variables s_i are independent, but they may have identical distributions.

2.3.6 Classification according to diversity

Another important classification relies on the diversity available at the receiver. In the simplest case, the diversity can be induced by the use of several antennas. With K antennas located sufficiently far from each other (for instance a half wavelength), the receive diversity is K. A one-source one-antenna transmission is referred to as Single Input Single Output (SISO) channel. For a larger number of sources or receive antennas, one shall talk about Multiple Input or Multiple Output channels (MIMO), respectively. This yields the classification of table 1.

In a more general framework, it is important to stress that there exist other forms of diversity than just spatial, and that they can be exploited even in the presence of a single antenna, giving rise to apparent SIMO or MIMO channels.

In the remainder and unless otherwise specified, the second hypothesis will be assumed, namely that sources are mutually statistically independent.

# Sources	# Sensors	
	1	K
1	**SISO**	**SIMO**
P	**MISO**	**MIMO**

Table 1: Classification according to transmit and receive diversities

2.3.7 Trivial filters

Define the set $\mathcal{T}$ of filters that do not affect the statistical independence. Then it is clear that this set contains the so-called *trivial filters* of the form:

$$\boldsymbol{G}[z] = \boldsymbol{P}\,\boldsymbol{D}[z]$$

where $\boldsymbol{P}$ is a permutation matrix, and $\boldsymbol{D}[z]$ a diagonal invertible filter.

Another simple way of understanding that is to notice that if $(\boldsymbol{H}, \boldsymbol{s})$ is solution, then so is $\big(\boldsymbol{H}\,(\boldsymbol{P}\boldsymbol{D})^{-1},\ \boldsymbol{P}\boldsymbol{D} \star \boldsymbol{s}\big)$ because statistical independence is not affected by scalar filtering nor permutation.

As a consequence, our goal is not to find a pair of solution $(\boldsymbol{H}, \boldsymbol{s})$, but a whole *equivalence class of solutions*, the equivalence relation being precisely defined as:

$$\boldsymbol{G} \sim \boldsymbol{H} \ \Leftrightarrow\ \exists \boldsymbol{T} \in \mathcal{T} : \boldsymbol{G} = \boldsymbol{T}\boldsymbol{H}$$

2.4 Complementary bibliographical remarks

This survey does not aim at being exhaustive. Hundreds of references have been unavoidably omitted for reasons of space, but many additional pointers can be found in [117], [90], or [44] for more recent pointers.

In the SISO framework, only the dynamic model is meaningful. The SISO Blind Equalization problem has been addressed as early as 1975, with the works of Sato [124], and later on Godard with his famous *Constant Modulus Algorithm* (CMA) [67], Benveniste [12], Donoho [57], Treichler [142], Shalvi [125]. In these approaches, the optimization criterion was either the *kurtosis*, which is a measure of deviation from normality (*i.e.* Gaussianity), or the Constant Modulus (*i.e.* on measures the deviation of the modulus of the output complex random variable from a constant). Another approach has been developed independently, and is more well known in astronomy. It is based on the use of bispectral statistics, as originally proposed by Marron [111], Matsuoka [112], Le Roux [101], or others [72] [64] [73] [97] [98].

Deterministic approaches exploiting the discrete character of the sources can be found in [105] [152] [104] [150] . Instead of source statistical independence, these approaches require some conditions of sufficient excitation.

In the SIMO case, *e.g.* when using several antennas in the presence of a single significant source, it is possible to resort to second-order statistics. This has been demonstrated by Slock [131] with oversampling diversity, Xu[151] , Moulines [116], Gesberg [66], among others. These algorithms are often referred to as *subspace based*, as they exhibit a subspace structure allowing to extract the source.

In the MIMO static case, the first contributions can be traced back to the eighties with Bar-Ness [9], Jutten [84], Comon [31]. Many subsequent works have addressed the static MIMO case since that time, and one can mention [33] for contrast-based approaches and *Jacobi* sweeping, [21] for approaches without pre-whitening, [23] for approximate joint diagonalization of matrices, [10] for entropy interpretations,

[52] for deflation, [26] [61] for the effect of nonstationarity on solutions based on sample estimates, [148] for an analytical Constant Modulus solution, [5] for the natural gradient formulation, [135] for non linear mixtures, and [6] for the geometry of Information. The case of discrete sources has been also addressed in [65] [137] [134] [149], among others.

In the MIMO dynamic case, the first contribution is apparently that of Comon in 1990 [32], with a Linear Prediction approach. Linear Prediction approaches have been followed independently six years later by Ding [54], and deeper analyses by Loubaton and co-authors [2] [69].

Subspace approaches for the MIMO case appear only in 1997 with the works of Loubaton and his students [68] [29], or [1]. Some identifiability issues by subspace techniques have been investigated in [107], based on properties of rational spaces [63]. See also Desbouvries [53].

Algebraic versions of the CMA have been investigated for MIMO mixtures in [149]. The use of the discrete character of the source distribution, or its constant modulus, have been also investigated in [141] [137] [70].

Contrast criteria have been proposed for MIMO dynamic mixtures in [37], and used for fixed step descent in [145]. Related references on Blind MIMO Equalization by kurtosis maximization include [130] [140].

The subject of Blind Identification (BI) has been comparatively much less studied. There are at least three motivations in going into BI. Firstly, it is useful as a pre-processing when trying to build a stable SISO equalizer, especially with Infinite Impulse Response; in fact, mainly FIR Blind Equalizers have been devised. Secondly, the BI problem cannot be avoided when dealing with Under-Determined mixtures (UDM), since they do not generally admit a linear inverse. And thirdly, in some problems, the quantity of interest is the mixture and not the sources themselves. In such a case, a direct BSS is inadequate.

Blind Identification is addressed in [12] [56] [73] [97] [152] for SISO channels, and in [139] [116] [146] [3] [1] for SIMO channels. BI of static MIMO mixtures is treated in [20] [4] [40] [59] [89] [13] [83]. BI of convolutive MIMO mixtures is eventually addressed in [34] [133] [138] [16] [107] [149] [55] [144] [106], among others.

General purpose books on *High-Order Statistics* (HOS) and their use in Blind Techniques include [117] [90] or [30]. Reference [82] is not a survey, but mainly dedicated to one adaptive algorithm, called FastICA, developed by the machine learning community; this book does not account correctly for previously developed works, nor provides any comparisons. Because of inconsistencies present in [82], we rather recommend [30] or [44].

2.5 Applications

Known application areas of Blind Techniques, are numerous, due to the general purpose character of the tool subsequently called *Independent Component Analysis* (ICA).

First in Antenna Array Processing, when antennas are ill calibrated (or when calibration varies too much with outside conditions such as temperature or pressure), as in Sonar, Automobile, Control (*e.g.* nuclear plants), or Biomedical.

In Surveillance and Interception (especially in the HF range), it is also useful to separate and classify unknown sources. For instance, Communications Intelligence, *ComInt*, consists of intercepting and decoding private electromagnetic transmissions; it is today widely used by the American National Security Agency (NSA) in order to spy on any selected user. In some cases, the antenna is partly unknown; examples include sonar buoys dropped on the see surface, ill-calibrated array in cheap equipment, or decalibration because of extreme conditions, *e.g.* in submarines.

In small diversity cases, Blind Techniques can be useful, as in Secondary Radar (air traffic control), Mobile Radio Communications, or Acoustics. Typical examples can be found in Speech or Image Processing.

ICA can also be viewed as a general tool to decompose tensors ($K-$way arrays), central in Data and Factor Analysis. ICA can thus be used in Econometrics, Chemometrics, Theoretical Psychology, Pharmacology, Exploratory Analysis, Arithmetic Complexity, or Machine Learning... In fact, for Supervised Learning in large dimensions, ICA may be useful as a preprocessing, allowing to reduce the minimum number of samples in order to estimate probability density functions.

3 Algebraic tools

Several algebraic tools are useful in this course. We first review classical Linear Algebra factorizations, then less well known Filter decompositions. Last, we give some definitions of Multi-linear tools, which will be used for manipulating tensors.

3.1 Principal Component Analysis and Standardization

The operation of *Whitening* consists of filtering the data so that the outputs are uncorrelated and of unit variance. This operation is also frequently called *Standardization* in the context of Statistics, when Mixture and Whitening are represented by static filters (i.e. mere matrices).

Singular Value decomposition

It is known that every matrix $\boldsymbol{M}$ may be decomposed into:

$$\boldsymbol{M} = \boldsymbol{U}\,\boldsymbol{\Sigma}\,\boldsymbol{V}^{\mathsf{H}}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary, and $\boldsymbol{\Sigma}$ is positive real diagonal diagonale. Columns $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ of $\boldsymbol{U}$ and $\boldsymbol{V}$ are the left and right singular vectors, respectively:

$$\boldsymbol{M}\,\boldsymbol{v}_i = \boldsymbol{u}_i\,\sigma_i \quad \boldsymbol{M}^{\mathsf{H}}\boldsymbol{u}_i = \boldsymbol{v}_i\,\sigma_i$$

In addition, $\boldsymbol{u}_i$ are eigenvectors of $\boldsymbol{M}\boldsymbol{M}^{\mathsf{H}}$, and $\boldsymbol{v}_i$ those of $\boldsymbol{M}^{\mathsf{H}}\boldsymbol{M}$, associated with σ_i^2. This decomposition is referred to as the Singular Value decomposition (SVD).

Filter decomposition

From Signal theory, we know that any scalar filter $g[z]$ can be decomposed into an *all-pass filter* $u[z]$ and a *minimum phase* filter $\ell[z]$. In other words:

$$\gamma[z] = u[z]\,\ell[z], \quad u[1/z^*]\,u[z] = 1,\ \forall z \tag{5}$$

with $\ell[z]$ having all its roots inside the unit circle. Recall that an all-pass filter is *lossless*, and has a flat frequency response (only its phase varies). This result concerns scalar filters, and thus applies to the SISO case.

More generally in the MIMO case, any filter defined by its Impulse Response matrix $\boldsymbol{F}[z]$, can be decomposed into the product of a Triangular filter $\boldsymbol{L}[z]$ with minimum phase and a *para-unitary* filter $\boldsymbol{U}[z]$:

$$\boldsymbol{F}[z] = \boldsymbol{U}[z]\,\boldsymbol{L}[z], \quad \boldsymbol{U}[1/z^*]^{\mathsf{H}}\boldsymbol{U}[z] = \boldsymbol{I},\ \forall z \tag{6}$$

Recall that a square Minimum Phase matrix is such that $\det(\boldsymbol{L}[z])$ has all its roots inside the unit circle. See [86] [147] for further readings on matrices with rational entries in one variable.

There is a particular case of importance, namely the static MIMO. Then, from the above decomposition, one can easily retrieve the classical QR decomposition of matrices:

$$\boldsymbol{F} = \boldsymbol{U}\,\boldsymbol{L}, \quad \boldsymbol{U}^{\mathsf{H}}\boldsymbol{U} = \boldsymbol{I} \tag{7}$$

where $\boldsymbol{L}$ is triangular and $\boldsymbol{U}$ unitary.

These decompositions are not unique.

Time whitening

Consider a scalar second order stationary process $x[k]$, and its $z-$ transform $x[z]$. Its power spectrum is given by

$$\gamma_x[z] \stackrel{\text{def}}{=} \mathrm{E}\{x[z]\,x[1/z^*]^*\}$$

Now from (5), it is possible to decompose the power spectrum as

$$\exists \ell[z] \,/ \quad \ell[z]\,\ell[1/z^*]^* = \gamma_x[z]$$

where filter $\ell[z]$ is not unique, and defined up to an all-pass filter. Filter $1/\ell[z]$ is called a *whitening filter* since the power spectrum after filtering with $\ell[z]$ is flat (and actually equal to 1).

Spatial whitening

Consider now a multivariate random variable $\boldsymbol{x}$, zero-mean for the sake of simplicity, with a finite covariance matrix:

$$\boldsymbol{\Gamma}_x \stackrel{\text{def}}{=} \mathrm{E}\{\boldsymbol{x}\,\boldsymbol{x}^{\mathsf{H}}\}$$

Then, from (7),

$$\exists \boldsymbol{L} \,/ \quad \boldsymbol{L}\,\boldsymbol{L}^{\mathsf{H}} = \boldsymbol{\Gamma}_x$$

In other words, we have found a matrix $\boldsymbol{L}$, such that $\boldsymbol{L}^{-1}\boldsymbol{x}$ has a unit variance. The variable $\tilde{\boldsymbol{x}} \stackrel{\text{def}}{=} \boldsymbol{L}^{-1}\boldsymbol{x}$ is then referred to as a *standardized random variable*.

Note that $\boldsymbol{L}$ may not be invertible if the covariance $\boldsymbol{\Gamma}_x$ is not full rank. In such a case, there always exists a rectangular matrix $\boldsymbol{R}$ such that $\boldsymbol{R}\boldsymbol{\Gamma}_x\boldsymbol{R}^{\mathsf{H}} = \boldsymbol{I}$. the latter matrix can be built from the SVD of $\boldsymbol{L}$, or from the EVD of $\boldsymbol{\Gamma}_x$.

Space-time whitening

Consider now the most general case of a multivariate second order stationary random process, with power spectral matrix:

$$\boldsymbol{\Gamma}_x[z] \stackrel{\text{def}}{=} \mathrm{E}\{\boldsymbol{x}[z]\,\boldsymbol{x}[1/z^*]^{\mathsf{H}}\}$$

Then, from (6)

$$\exists \boldsymbol{L}[z] \,/ \quad \boldsymbol{L}[z]\,\boldsymbol{L}[1/z^*]^{\mathsf{H}} = \boldsymbol{\Gamma}_x[z]$$

thus, the filter with complex gain matrix $\boldsymbol{L}[z]$ may be used to build a whitening filter $\boldsymbol{R}[z]$, defined up to a para-unitary filter. If $\boldsymbol{L}[z]$ admits an inverse, then we may take $\boldsymbol{R}[z] = \boldsymbol{L}[z]^{-1}$, and $\tilde{\boldsymbol{x}}[k] = \boldsymbol{R} \star \boldsymbol{x}[k]$.

Source extraction

This suggests a Blind MIMO Equalization (or Source Separation) algorithm in *three stages*:

1. Whitening by a filter $\boldsymbol{L}$ (non unique).
2. Separation with a para-unitary filter $\boldsymbol{U}$ (unique modulo the equivalence class).
3. Construction of an extracting filter $\boldsymbol{W}$, which can be merely $\boldsymbol{UL}$, or based on an estimate of $\boldsymbol{H}$, e.g. $\boldsymbol{L}^{-1}\boldsymbol{U}^{\mathsf{H}}$. For instance the *Space-time Matched Filter* (SMF) takes the form:

$$\boldsymbol{W} = \boldsymbol{H}^{\mathsf{H}}\boldsymbol{R}_x^{-1}$$

whereas the Weighted Least Squares estimate is usable if in addition the noise spatial coherence $\boldsymbol{B}$ is known [42]:

$$\boldsymbol{W} = (\boldsymbol{H}^{\mathsf{H}}\boldsymbol{B}^{-1}\boldsymbol{H})^{-1}\boldsymbol{H}^{\mathsf{H}}\boldsymbol{B}^{-1}$$

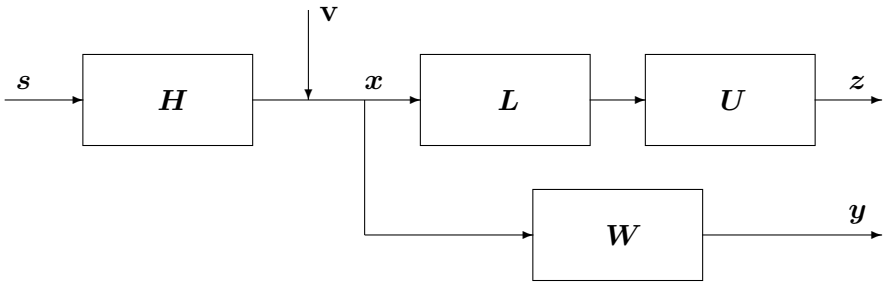


Figure 1: Processing line of Blind Source Separation with prewhitening

In practice, pre-whitening is only rarely chosen for dynamic mixtures because of difficulties of implementation, whereas it is widely used for static mixtures. Also note that if the noise covariance is known up to a multiplicative scalar (i.e. $\boldsymbol{B}$ is known), pre-whitening can be based on $\boldsymbol{B}$ instead of $\boldsymbol{R}_x$, leading to a model where additive noise is white.

3.2 Jacobi sweeping

In this section, we recall the standard technique devised by Jacobi for diagonalizing a symmetric matrix by successive applications of plane rotations. Define a so-called *Givens* plane rotation, $\boldsymbol{Q}$, equal to the identity matrix except for four entries: $G(i,i) = G(j,j) = \cos\theta$, $G(i,j) = -G(j,i) = \sin\theta$. The simultaneous application of $\boldsymbol{Q}$ on the left and on $\boldsymbol{Q}^{\mathsf{T}}$ on the right of any symmetric matrix $\boldsymbol{A}$ allows to set two zeros at the (i,j) and (j,i) coordinates:

$$\begin{pmatrix} c & . & s & . \\ . & 1 & . & . \\ -s & . & c & . \\ . & . & . & 1 \end{pmatrix} \boldsymbol{A} \begin{pmatrix} c & . & -s & . \\ . & 1 & . & . \\ s & . & c & . \\ . & . & . & 1 \end{pmatrix} = \begin{pmatrix} X & x & 0 & x \\ x & . & x & . \\ 0 & x & X & x \\ x & . & x & . \end{pmatrix}$$

Entries denoted with a dot are unchanged by this transformation, where as the others are modified. More precisely, entries depicted by 'X' are maximized. Sweeping successively all the pairs allows to monotically increase the sum of squares of all diagonal terms, or, equivalently (because the whole

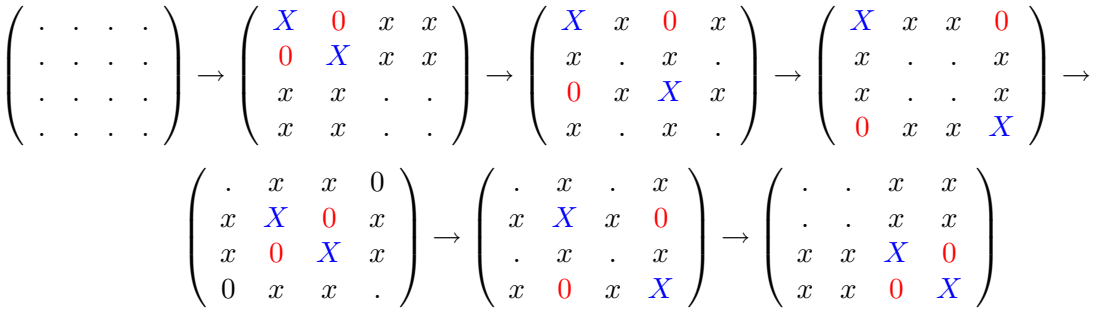


Figure 2: cyclic by rows/columns algorithm for a 4×4 real symmetric matrix. X: maximized, x: minimized, 0: canceled, $\cdot$: unchanged.

Frobenius norm of matrix $\boldsymbol{Q}\,\boldsymbol{A}\,\boldsymbol{Q}^\mathsf{T}$ remains unchanged), to monotically decrease all the extra-diagonal terms. An analysis of stationary points shows that such an algorithm sweeping all the pairs in a prescribed order, as depicted in figure 2, converges to a diagonal matrix. Proofs are given in algebra textbooks, but a compact proof may be found in [36].

3.3 Tensor operations

We have just surveyed some basic decompositions of matrices. Matrices are just two-way arrays, that is, arrays whose entries are described by two indices. In the remainder, we shall need some simple operators able to act on arrays with more than two ways. in the framework of signal processing, such arrays often result from the calculation of statistics, and hence enjoy multi-linearity properties allowing them to deserve the name of tensors. The number of indices (*i.e.* ways) is sometimes called the *tensor order*.

First, the *outer product* allows to increase the number of ways of an array. For instance, the outer product of two column vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ is the rank-one matrix $\boldsymbol{A} = \boldsymbol{u}\,\boldsymbol{v}^\mathsf{T}$, which is a two-way array whose entries are $A_{ij} u_i\, v_j$. More generally, given two tensors of entries $A_{i..j}$ and $B_{k..\ell}$, of order m and n respectively, one can build a tensor of order $m+n$ by defining

$$\boldsymbol{C} \stackrel{\text{def}}{=} \boldsymbol{A} \circ \boldsymbol{B}, \quad C_{i..jk..\ell} = A_{i..j} B_{k..\ell} \tag{8}$$

Another useful operation on tensor is the *contraction*. Given two tensors $A_{ij..k}$ and $B_{ip..q}$, of same first dimension, the contraction on the first index is defined by

$$\boldsymbol{C} \stackrel{\text{def}}{=} \boldsymbol{A} \underset{1}{\bullet} \boldsymbol{B}, \quad C_{j..kp..q} = \sum_i A_{ij..k} B_{ip..q} \tag{9}$$

More generally, the contraction product between two tensors on the kth index will be denoted by $\bullet_k$, and is possible if their kth dimension is the same.

As an example, the matrix-vector product can be rewritten in the form of a contraction: $\boldsymbol{A}\,\boldsymbol{v} = \boldsymbol{A}^\mathsf{T} \bullet_1 \boldsymbol{v}$. Similarly, the scalar product between two vectors is the contraction: $\boldsymbol{u}^\mathsf{T}\boldsymbol{v} = \boldsymbol{u} \bullet_1 \boldsymbol{v}$.

The *Kronecker product* between two tensors of same order is denoted by $\boldsymbol{A} \otimes \boldsymbol{B}$, and consists of a tensor formed by the blocks $A_{ij..k}\boldsymbol{B}$. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are identical, then it is clear that $\boldsymbol{A} \otimes \boldsymbol{A}$ is redundant. Therefore, it may be useful to define a non redundant array $\boldsymbol{A} \oslash \boldsymbol{A}$. For instance for a vector $\boldsymbol{v}$ of dimension n, the vector $\boldsymbol{v} \oslash \boldsymbol{v}$ is of dimension $n(n+1)/2$.

The column-wise Kronecker product between two matrices is often called the *Khatri-Rao product*, and is denoted $\boldsymbol{A} \odot \boldsymbol{B}$.

Now there exists a mapping $\mathbf{vec}\{\}$ that maps a tensor to a vector. For instance, for a 3−way tensor $\boldsymbol{A}$ of dimension $m \times n \times p$, $\mathbf{vec}\{\boldsymbol{A}\}$ is of dimension mnp. If this tensor is square completely symmetric, that is if $n = m = p$ and $A_{ijk} = A_{\sigma(i,j,k)}$ for all indices (i, j, k) and any permutation σ, then it is possible to map $\boldsymbol{A}$ to a lower dimensional vector, $\mathbf{vecs}\{\boldsymbol{A}\}$ of dimension $n(n+1)(n+2)/6$. By convention, the entries appearing k times in tensor $\boldsymbol{A}$, appear in vector $\mathbf{vecs}\{\boldsymbol{A}\}$ with a weighting $\sqrt{k}$; this has the advantage that the operator $\mathbf{vecs}\{\cdot\}$ preserves the L^2 norm.

4 Statistical tools

First, definitions and properties are given for real scalar random variables in sections 4.1 and 4.2. In a second stage, these statements are extended to multivariate real random variables in section 4.2.2. The case of complex variables is addressed in sections 4.4 and 4.4.2. These tools will allow to introduce various definitions of statistical independence in section 5.

4.1 Characteristic functions

For a real random variable x, let $F_x(u)$ denote its cumulative probability distribution, which corresponds to the probability that $x < u$. When it exists, denote its density $p_x(u)$; we have that $dF(u) = p_x(u)\, du$.

One defines *generalized moments* as

$$\mathrm{E}\{g(x)\} = \int_{-\infty}^{+\infty} g(u)\, dF(u)$$

In practice, $g(u)$ is often a monomial, which yields the usual *non central moments*:

$$\mu'_{x\,(k)} = \mathrm{E}\{x^k\} \tag{10}$$

If the exponential function is used instead, one gets the *first characteristic function*:

$$\Phi(t) = \int e^{\jmath\, tu}\, dF(u) \tag{11}$$

where $\jmath$ is the square root of -1.

Example 1: Real Gaussian variable. For a *real Gaussian variable*, the first characteristic function takes the form

$$\Phi(t) = \exp[\jmath \mu'_1\, t - \frac{1}{2}\, \mu_2\, t^2] \tag{12}$$

For multivariate random variables, the definition is the same, but the product tu is replaced by the scalar product $\boldsymbol{t}^{\mathsf{T}}\, \boldsymbol{u}$:

$$\Phi(\boldsymbol{t}) = \int e^{\jmath\, \boldsymbol{t}^{\mathsf{T}} \boldsymbol{u}}\, dF(\boldsymbol{u})$$

The first characteristic function obviously satisfies the following properties:

$$\Phi(0) = 1, \quad |\Phi(t)| \leq 1, \forall t$$

One can also show that Φ is uniformly continuous. When the density $p_x(\boldsymbol{u})$ exists, then $\Phi(\boldsymbol{t})$ can be viewed as its Fourier transform.

Another interesting property (difficult to prove) is the Lebesgue decomposition [109]:

$$\Phi = a_d\,\Phi_d + a_{ac}\,\Phi_{ac} + a_s\,\Phi_s$$

where Φ_d is the characteristic function of a discrete distribution, Φ_{ac} of an absolutely continuous distribution, and Φ_s of a singular distribution, and a_d, a_{ac} and a_s are probabilities that sum up to 1. The two first distributions are well known: Φ_d is almost periodic, and Φ_{ac} tends to zero at infinity. But the third term is somewhat abstract, and not really encountered in signal processing.

Because $\Phi(\boldsymbol{t})$ is non zero in a neighborhood of the origin, one can define its logarithm, the *second characteristic function*:

$$\Psi(\boldsymbol{t}) = \mathrm{Log}\Phi(\boldsymbol{t}) \tag{13}$$

This function always exists in a (sufficiently small) neighborhood of the origin, and is unique as long as $\Phi \neq 0$.

4.2 Moments and cumulants

4.2.1 Scalar random variables

If Φ can be expanded in Taylor series about the origin, then its coefficients are related to *moments*:

$$\mu'_{x\,(r)} \stackrel{\text{def}}{=} \mathrm{E}\{X^r\} = (-\jmath)^r \left.\frac{\partial^r \Phi(t)}{\partial t^r}\right|_{t=0} \tag{14}$$

It is usual to introduce *central moments* $\mu_{x\,(r)}$ as the moments of the centered variable $x - \mu'_{x\,(1)}$.

Similarly, if Ψ may be expanded in Taylor series about the origin, then its coefficients are the *cumulants*:

$$\mathcal{C}_{x\,(r)} \stackrel{\text{def}}{=} \mathrm{Cum}\{\underbrace{X, X, \ldots, X}_{r\ times}\} = (-\jmath)^r \left.\frac{\partial^r \Psi(t)}{\partial t^r}\right|_{t=0} \tag{15}$$

The relation between moments and cumulants can be obtained by expanding the logarithm and grouping terms of same order together.

Example 2: Cumulants may not exist. Note that expansion (14) may not always exist. The classical counter-example is that of the Cauchy distribution

$$p(u) = \frac{1}{\pi\,(1+u^2)}$$

for which moments and cumulants of order larger than 1 are infinite.

Example 3: Cumulants of order 2, 3 and 4. The cumulant of 2nd order, κ_2, is nothing but the variance: $\mu'_{(2)} - {\mu'_{(1)}}^2 = \mathcal{C}_{(2)}$. And for zero-mean random variables, cumulants of order 3 and 4 are related to moments by: $\kappa_3 = \mu_3$ and $\kappa_4 = \mu_4 - 3\,{\mu_2}^2$.

Example 4: Skewness and Kurtosis. The *skewness* is a 3rd order normalized cumulant: $\mathcal{K}_3 \stackrel{\text{def}}{=} \kappa_3/{\kappa_2}^{3/2}$. The *kurtosis* is a normalized 4th order cumulant $\mathcal{K}_4 \stackrel{\text{def}}{=} \kappa_4/{\kappa_2}^2$

It is clear that skewness and kurtosis are null for any Gaussian random variable. These quantities can serve as measures of deviation from Gaussianity. In fact, random variables having a negative (resp. positive) kurtosis can be called *platykurtic* (resp. *leptokurtic*) [87]. Conversely, random variables have a zero kurtosis (referred to as *mesokurtic*) are not necessarily Gaussian.

It is sometimes spoken in the literature of *sub-Gaussian* or *super-Gaussian* random variables. It is important to stress that this feature is not well defined, and that at least three different definitions can be found.

First, Benveniste [12, page 390] proposes a definition involving the monotonicity of

$$f(u) = -\frac{1}{u}\frac{d\,\log p_x(u)}{du}.$$

When $f(u)$ is strictly increasing (resp. decreasing), $p_x(u)$ is said super-Gaussian (resp. sub-Gaussian). It is clear that some distributions are neither one or the other.

Second, some authors [154] linked the super-Gaussian (resp. sub-Gaussian) character to the fact that distributions tails are above (resp. below) those of the Gaussian with same mean and variance. This definition coincides with that of *platykurtic* (resp. *leptokurtic*) when the even part of the density intersects twice the Gaussian of same mean and variance [110]. In other words, the definitions are not always the same.

xx give example

4.2.2 Multivariate random variables

Denote the cumulants $\kappa_{ij..\ell} = \mathrm{Cum}\{X_i, X_j, ...X_\ell\}$. As explained above, expressions of moments as a function of cumulants can be obtained by expanding the logarithm and grouping terms of same order together. This yields for instance:

$$\begin{aligned} \mu'_i &= \kappa_i \\ \mu'_{ij} &= \kappa_{ij} + \kappa_i\,\kappa_j \\ \mu'_{ijk} &= \kappa_{ijk} + [3]\,\kappa_i\,\kappa_{jk} + \kappa_i\,\kappa_j\,\kappa_k \end{aligned} \tag{16}$$

In the relation above, we have used Mccullagh's *bracket notation* [113] defined below.

A sum of k terms that can be deduced from each other by permutation of indices is denoted by the number k between brackets followed by a single monomial describing the generic term. Simple examples will do it better than a long explanation.

$$[3]\,\delta_{ij}\,\delta_{kl} = \delta_{ij}\,\delta_{kl} + \delta_{ik}\,\delta_{jl} + \delta_{il}\,\delta_{jk}, \tag{17}$$

$$[3]\,a_{ij}\,b_k\,c_{ijk} = a_{ij}\,b_k\,c_{ijk} + a_{ik}\,b_j\,c_{ijk} + a_{jk}\,b_i\,c_{ijk}. \tag{18}$$

The presence of the bracket yields an implicit summation; all terms with r indices are completely symmetric order-r tensors. The number of distinct monomials that may be obtained by permutation is equal to the integer appearing between brackets. As additional examples, the following expressions are consistent:

$$[3]\,a_i\delta_{jk}, \quad [6]\,a_i\,a_j\,\delta_{kl}, \quad [10]\,b_i\,b_j\,b_k\,\delta_{lm}, \quad [35]\,A_{ijk}\,B_{abcd}\,C_{ijkabcd}.$$

Relations (16) can be inverted in order to obtain cumulants as a function of moments. In the case of non central random variables, *multivariate cumulants* of order 3 and 4 can be given in a compact form as a function of *multivariate moments* as:

$$\mathcal{C}_{ijk} = \mu'_{ijk} - [3]\,\mu'_i\,\mu'_{jk} + 2\mu'_i\,\mu'_j\,\mu'_k, \tag{19}$$

$$\begin{aligned} \mathcal{C}_{ijkl} &= \mu'_{ijkl} - [4]\,\mu'_i\,\mu'_{jkl} - [3]\mu'_{ij}\,\mu'_{kl} \\ &+ 2\,[6]\,\mu'_i\,\mu'_j\,\mu'_{kl} - 6\mu'_i\,\mu'_j\,\mu'_k\,\mu'_l. \end{aligned} \tag{20}$$

On the other hand, if variables are all zero-mean, then they simplify into:

$$\begin{aligned} \mathcal{C}_{ij} &= \mu_{ij}, & (21)\\ \mathcal{C}_{ijk} &= \mu_{ijk}, & (22)\\ \mathcal{C}_{ijkl} &= \mu_{ijkl} - [3]\,\mu_{ij}\,\mu_{kl}. & (23) \end{aligned}$$

and for orders 5 and 6:

$$\begin{aligned} \mathcal{C}_{ijklm} &= \mu_{ijklm} - [10]\,\mu_{ij}\,\mu_{klm},\\ \mathcal{C}_{ijklmn} &= \mu_{ijklmn} - [15]\,\mu_{ij}\,\mu_{klmn}\\ &- [10]\,\mu_{ijk}\,\mu_{lmn} + 2\,[15]\,\mu_{ij}\,\mu_{kl}\,\mu_{mn}. \end{aligned}$$

4.2.3 Properties of cumulants

Cumulants enjoy several useful properties. Some of them are shared by moments, but others are not. First of all, moments and cumulants enjoy the *multi-linearity property*:

$$\begin{aligned} \mathrm{Cum}\{\alpha X,\, Y, .., Z\} &= \alpha\,\mathrm{Cum}\{X,\, Y, .., Z\} & (24)\\ \mathrm{Cum}\{X_1 + X_2,\, Y, .., Z\} &= \mathrm{Cum}\{X_1,\, Y, .., Z\} + \mathrm{Cum}\{X_2,\, Y, .., Z\} \end{aligned}$$

Another obvious property directly results from the definition, namely that of invariance by permutation of indices

$$\mathrm{Cum}\{X_1,\, X_2,\, ..X_r\} = \mathrm{Cum}\{X_{\sigma(1)},\, X_{\sigma(2)},\, ..X_{\sigma(r)}\}$$

In other words, rth order Cumulants (or Moments) can be stored in $r-$way arrays that deserve the name of rth order *tensors*.

Let's now turn to properties that are specific to cumulants. First, they are invariant to translation; this means that $\forall r > 1$ and $\forall h$ constant:

$$\mathrm{Cum}\{X_1 + h,\, X_2, .., X_r\} = \mathrm{Cum}\{X_1,\, X_2, .., X_r\},$$

This property is sometimes referred to as the *shift invariance of cumulants*. Next, cumulants of a set of random variables are null as soon as this set can be split into two statistically independent subsets:

$$\{X_1, \ldots, X_p\} \text{ independent of } \{Y_1, \ldots, Y_q\} \Rightarrow \mathrm{Cum}\{X_1, \ldots, X_p, Y_1, \ldots, Y_q\} = 0$$

A consequence of this property is the *additivity of cumulants*:

$$\begin{aligned} \mathrm{Cum}\{X_1 + Y_1,\, X_2 + Y_2, .., X_r + Y_r\} &= \mathrm{Cum}\{X_1,\, X_2, .., X_r\} & (25)\\ &+ \mathrm{Cum}\{Y_1,\, Y_2, .., Y_r\} \end{aligned}$$

whenever $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent.

4.3 The cumulant-based Central Limit Theorem

The *Central Limit theorem* can be revisited (in a weaker form) with the help of cumulants. Let $x[n], 1 \leq n \leq N$ be N statistically independent random variables, each having a bounded cumulant of order r, denoted $\mathcal{C}_{x[n]\,(r)} = \kappa(r;n)$. Define the average cumulant of order r

$$\bar{\kappa}(r) = \frac{1}{N}\sum_{n=1}^{N} \kappa(r;n)$$

and the average random variable

$$y = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} (x(n) - \bar{\kappa}(1;n))$$

Then, if $\bar{\kappa}(r)$ remain bounded, the probability distribution f_y tends to a Gaussian law as $N \to \infty$.

Proof. Because of the mutual independence of variables $x[n]$ and by multi-linearity of cumulants, cumulants of order r of $\sum_n x[n]$ are $\sum_n \kappa(r;n)$. Hence, cumulants of order r of y are $\mathcal{C}_{y\,(r)} = \frac{\bar{\kappa}(r)}{N^{r/2-1}}$, $\forall r \geq 2$. It is thus clear that all these cumulants tend to zero as N tends to infinity, because $\bar{\kappa}(r)$ are bounded. Next, the second characteristic function, $\Psi_y(u)$, can be expanded in Taylor series because all the cumulants exist and are finite. In this expansion of $\Psi_y(u)$, except for the second term $\mathcal{C}_{y\,(2)} = \bar{\kappa}(2)$, the infinite sum of all remaining terms also tends to zero, for any fixed u. This shows that $\Psi_y(u)$ tends to the characteristic function of a zero-mean Gaussian variable, and hence the convergence in distribution, provided the limit of $\bar{\kappa}(2)$ is positive, and provided the limiting function is continuous at the origin. □

4.4 Complex variables

4.4.1 Definitions

In Signal Processing, it is often suitable to manipulate complex valued random variables, essentially for the sake of simplicity in the notation. Therefore, it is necessary to introduce the relevant definitions. Let $\boldsymbol{z} = \boldsymbol{x} + \jmath\boldsymbol{y}$ be a *complex random variable*, where $\boldsymbol{x}$, and $\boldsymbol{y}$ are real multivariate random variables of same dimension. Also define the deterministic complex variable $\boldsymbol{w} = \boldsymbol{u} + \jmath\boldsymbol{v}$, where $\boldsymbol{u}$ and $\boldsymbol{v}$ are mutivariate real variables of same dimension. The probability distribution of $\boldsymbol{z}$ is defined as the joint distribution of $(\boldsymbol{x}, \boldsymbol{y})$:

$$f_{\boldsymbol{z}}(\boldsymbol{w}) \stackrel{\text{def}}{=} f_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{u}, \boldsymbol{v})$$

This yields without difficulty the definitions below, admitting a compact writing. The cumulative distribution takes the form:

$$F_{\boldsymbol{z}}(\boldsymbol{w}) \stackrel{\text{def}}{=} F_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{u}, \boldsymbol{v})$$

and the first characteristic function

$$\Phi_{\boldsymbol{z}}(\boldsymbol{w}) = \mathrm{E}\{e^{\Re[\boldsymbol{z}^{\mathsf{H}}\boldsymbol{w}]}\} \tag{26}$$

since $\Re[\boldsymbol{z}^{\mathsf{H}}\boldsymbol{w}] = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{u} + \boldsymbol{y}^{\mathsf{T}}\boldsymbol{v}$

Now contrary to real variables, there is not a unique way of defining a cumulant (or a moment) of order r of a complex random variable; in fact, it depends on the number of conjugated terms. It is thus necessary to be able to distinguish between complex random variables that are conjugated and those that are not. For this purpose, one introduces a specific notation, with superscripts:

$$\mathrm{Cum}\{X_i, ..X_j, X_k{}^*, ..X_\ell{}^*\} \stackrel{\text{def}}{=} \mathcal{C}_{i..j}^{k..\ell} \tag{27}$$

Example 5: The covariance matrix. The covariance matrix of a complex random variable is

$$\mathrm{E}\{z_i\, z_j{}^*\} - \mathrm{E}\{z_i\}\,\mathrm{E}\{z_j\}^* = \mathcal{C}_{z\,i}^{\ \ j}$$

On the other hand, as will be defined in the next section, if the random variable $\boldsymbol{z}$ is *circular at order 2*, then the moment $\mu_{x\,i}^{\ \ j} = \mathrm{E}\{\boldsymbol{z}\,\boldsymbol{z}^{\mathsf{T}}\}_{ij} = \mu_{x\,i}^{\ \ j}$ is null for all (i, j).

4.4.2 Circularity

A complex random variable z is *circular in the strict sense* if its distribution does not depend on the phase of z. For a multivariate complex random variable $\boldsymbol{z}$, strict sense stationarity means that

$$\boldsymbol{z} \text{ and } \boldsymbol{z}\, e^{\jmath\theta},\ \forall \theta \in \mathbb{R}$$

have the same joint distribution, and hence statistical properties at any order.

A weaker definition of circularity is often sufficient. According to definition (27), there may be up to $r!/\lfloor r/2 \rfloor!$ distinct definitions, depending on what variables are conjugated. Among all these definitions, only one will be called *circular cumulant*, namely the one having exactly half of its arguments conjugated. All other cumulants may be called *non circular* cumulants. Note that there exist circular cumulants only at even orders.

For instance, the cumulant below is circular

$$\mathcal{C}_{\boldsymbol{z}\,ij}^{\;k\ell} = \mathrm{Cum}\{z_i,\, z_j,\, {z_k}^*,\, {z_\ell}^*\}$$

whereas these ones are non circular

$$\mathcal{C}_{\boldsymbol{z}\,ijl}^{\;\ell} = \mathrm{Cum}\{z_i,\, z_j,\, z_k,\, {z_\ell}^*\}$$
$$\mathcal{C}_{\boldsymbol{z}\,ijk\ell} = \mathrm{Cum}\{z_i,\, z_j,\, z_k,\, z_\ell\}$$

A multivariate complex random variable $\boldsymbol{z}$ is said to be *circular at order* r if its non circular cumulants of order r are all null:

$$p \neq r - p \Rightarrow \mathrm{Cum}\{z_1, .., z_p, {z_{p+1}}^*, .., {z_r}^*\} = 0 \tag{28}$$

4.5 Examples

Example 6: PSK random variables. For a PSK-4 random variable, $ZZ^* = 1$ and consequently:

$$\mathcal{C}_{(2)} = \mathrm{E}\{Z^2\} = 0, \mathcal{C}_{(2)}^{(2)} = -1, \mu_{(4)}^{(0)} = 1, \mathcal{C}_{(4)}^{(0)} = 1$$

It is thus circular up to order 3.

Example 7: Uniform distribution. A random variable uniformly distributed in the interval $[-u, +u]$, has the following expressions for moments and cumulants: $\mu_{2k} = \frac{u^{2k}}{2k+1}, \mathcal{C}_4 = \frac{u^4}{5} - 3\,\frac{u^4}{9} = -2\,\frac{u^4}{15}$, $\mathcal{K}_4 = -\frac{6}{5}$.

Example 8: Binary distribution. Let x be a random variable taking two values x_1 and x_2 with probabilities P_1 and P_2. We must have $P_1 = \frac{1}{1+a^2}$, $P_2 = \frac{a^2}{1+a^2}$, and $x_1 = -a$, $x_2 = 1/a$, for some free parameter a, if x is zero-mean and unit variance. Then the skewness is $\mathcal{C}_3 = 1/a - a$ and the kurtosis is $\mathcal{C}_4 = P_2\,[a^2 + 1/a^4] - 3$. As depicted in figure 3, the minimal value achieved by the kurtosis is -2.

Example 9: Two sources without noise. Assume a static model $\boldsymbol{x} = \boldsymbol{H}\,\boldsymbol{s}$ of dimension 2, where $\boldsymbol{s}$ has two statistically independent components. Also assume that a spatial pre-whitening has been performed, so that the standardized observation exactly follows the model below:

$$\tilde{\boldsymbol{x}} = \begin{pmatrix} \cos\alpha & -\sin\alpha\, e^{\jmath\varphi} \\ \sin\alpha\, e^{-\jmath\varphi} & \cos\alpha \end{pmatrix} \boldsymbol{s} \tag{29}$$

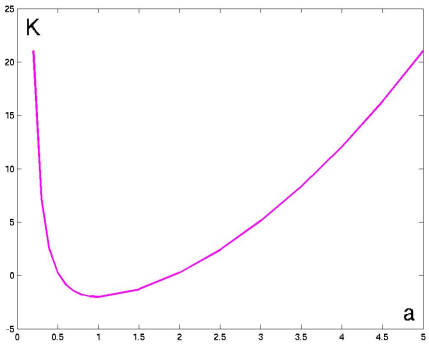


Figure 3: Kurtosis of a zero-mean standardized binary random variable

Denote $\gamma_{ij}^{k\ell} = \mathrm{Cum}\{\tilde{x}_i, \tilde{x}_j, \tilde{x}_k{}^*, \tilde{x}_\ell{}^*\}$ the fourth order circular cumulants of $\tilde{\boldsymbol{x}}$ and $\kappa_i = \mathrm{Cum}\{s_i, s_i, s_i{}^*, s_i{}^*\}$ those of s_i. Using the multi-linearity property of cumulants, one gets the input-output relations:

$$\begin{array}{rcl} \gamma_{12}^{12} &=& \cos^2\alpha\,\sin^2\alpha\,(\kappa_1+\kappa_2) \\ \gamma_{11}^{12} &=& \cos^3\alpha\,\sin\alpha\,e^{\jmath\varphi}\,\kappa_1 - \cos\alpha\,\sin^3\alpha\,e^{\jmath\varphi}\,\kappa_2 \\ \gamma_{12}^{22} &=& \cos\alpha\,\sin^3\alpha\,e^{\jmath\varphi}\,\kappa_1 - \cos^3\alpha\,\sin\alpha\,e^{\jmath\varphi}\,\kappa_2 \end{array} \tag{30}$$

From above, one deduces that $\gamma_{12}^{22} - \gamma_{11}^{12} = -2\,\gamma_{12}^{12}\,\cot 2\alpha\,e^{\jmath\varphi}$ This suggests a simple blind separation algorithm.

Example 10: Gauss distribution. Let x be a real Gaussian random variable of probability distribution $p_x(u) = \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{u^2}{2\,\sigma^2})$. Then even order moments are given by

$$\mu_{x\,(2r)} = \sigma^{2r}\,\frac{(2r)!}{r!\,2^r}. \tag{31}$$

4.6 Statistical Independence and geometry of information

The purpose of this section is to justify the use of the Mutual Information as the appropriate measure of statistical independence, and to decompose it into the appropriate elementary terms.

Definition 1 *Let $\boldsymbol{x}$ be a random vector of size K, admitting a density $p_{\boldsymbol{x}}(\boldsymbol{u})$. Components x_i of $\boldsymbol{x}$ are said to be mutually statistically independent if and only if the joint distribution equals the product of the marginal distributions:*

$$p_{\boldsymbol{x}}(\boldsymbol{u}) = \prod_{i=1}^{N} p_{x_i}(u_i). \tag{32}$$

Thus a quite natural way of measuring statistical independence between random variables x_i is to measure the distance $\delta(p_{\boldsymbol{x}}, \prod_i p_{x_i})$ between these distributions. Among the available distance measures, one is quite often used, namely the *Kullback divergence*:

$$\mathrm{K}(p_{\boldsymbol{x}}, p_{\boldsymbol{y}}) \stackrel{\mathrm{def}}{=} \int p_{\boldsymbol{x}}(\boldsymbol{u})\,\log\frac{p_{\boldsymbol{x}}(\boldsymbol{u})}{p_{\boldsymbol{y}}(\boldsymbol{u})}\,d\boldsymbol{u}. \tag{33}$$

Note the word *divergence* has been used, since this distance measure is not a symmetric in both its arguments (hence it doesn't deserve the name of *distance*).

Proposition 2 *The Kullback divergence is always positive and cancels if and only if both distributions are equal almost everywhere (that is except on a set of null measure):*

$$\mathrm{K}(p_{\boldsymbol{x}}, p_{\boldsymbol{y}}) = 0 \Leftrightarrow p_{\boldsymbol{x}}(u) \stackrel{\text{ae}}{=} p_{\boldsymbol{y}}(u). \tag{34}$$

Proof. For every positive real number w, we have the convexity inequality $\log w \leq w - 1$, with equality if and only if $w = 1$. By applying this inequality to the ratio $p_{\boldsymbol{x}}(\boldsymbol{u})/p_{\boldsymbol{y}}(\boldsymbol{u})$, one gets:

$$-\mathrm{K}(p_{\boldsymbol{x}}, p_{\boldsymbol{y}}) \leq \int p_{\boldsymbol{x}}(\boldsymbol{u}) \, [\frac{p_{\boldsymbol{y}}(\boldsymbol{u})}{p_{\boldsymbol{x}}(\boldsymbol{u})} - 1] \, d\boldsymbol{u}.$$

Yet, the right hand side is always null because $\int p(\boldsymbol{u}) \, \boldsymbol{u} = 1$ for any probability density. On the other hand, the function $\log w$ being tangent to $w - 1$ at $w = 1$, equality holds only if $p_{\boldsymbol{y}}(\boldsymbol{u})/p_{\boldsymbol{x}}(\boldsymbol{u}) = 1$ for almost all $\boldsymbol{u}$. □

Proposition 3 *The Kullback divergence is invariant under invertible transform of both its arguments.*

Proof. Let $\boldsymbol{Y} = A\boldsymbol{y}$ and $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{A}$ is an invertible matrix. Then $p_X(\boldsymbol{v}) = p_{\boldsymbol{x}}(\boldsymbol{A}^{-1}\boldsymbol{v})/|\det(\boldsymbol{A})|$, and $p_Y(\boldsymbol{v}) = p_{\boldsymbol{y}}(\boldsymbol{A}^{-1}\boldsymbol{v})/|\det(\boldsymbol{A})|$. This yields then

$$\mathrm{K}(p_X, p_Y) = \int p_{\boldsymbol{x}}(\boldsymbol{A}^{-1}\boldsymbol{v}) \, \log \frac{p_{\boldsymbol{x}}(\boldsymbol{A}^{-1}\boldsymbol{v})}{p_{\boldsymbol{y}}(\boldsymbol{A}^{-1}\boldsymbol{v})} \, \frac{d\boldsymbol{v}}{|\det \boldsymbol{A}|}.$$

define $\boldsymbol{u} = \boldsymbol{A}^{-1}\boldsymbol{v}$, that is $d\boldsymbol{v} = |\det \boldsymbol{A}| d\boldsymbol{u}$. Then

$$\mathrm{K}(p_X, p_Y) = \int p_{\boldsymbol{x}}(\boldsymbol{u}) \, \log \frac{p_{\boldsymbol{x}}(\boldsymbol{u})}{p_{\boldsymbol{y}}(\boldsymbol{u})} \, d\boldsymbol{u},$$

which terminates the proof. □

Now the Kullback divergence applied to $p_{\boldsymbol{y}}(\boldsymbol{u}) \stackrel{\text{def}}{=} \prod p_{x_i}(u_i)$ leads to the *Mutual Information* (MI), as a measure of independence:

$$I(p_{\boldsymbol{x}}) = \int p_{\boldsymbol{x}}(\boldsymbol{u}) \, \log \frac{p_{\boldsymbol{x}}(\boldsymbol{u})}{\prod_{i=1}^{N} p_{x_i}(u_i)} \, d\boldsymbol{u}. \tag{35}$$

This quantity is known as the average mutual information in coding theory. Because of proposition 2, it is always positive, and cancels if and only if random variables x_i are mutually independent. The MI is thus a first possible contrast function [35] [118].

Contrary to what one could perhaps guess, the MI is not invariant by invertible change of coordinates, although the Kullback divergence was. The reason is that a product of marginals is not generally transformed into another product of marginals. To check this, just consider the following counter-example.

Example 11: MI of a Gaussian variate. Take a real Gaussian random variable of size K, of density

$$g_{\boldsymbol{x}}(\boldsymbol{u}) = [2\pi]^{-K/2} \, [\det \boldsymbol{V}]^{-1/2} \, e^{-\frac{1}{2}\boldsymbol{u}^{\intercal}\boldsymbol{V}^{-1}\boldsymbol{u}}, \tag{36}$$

where $\boldsymbol{V}$ is an invertible covariance matrix. Then its Mutual Information is given by

$$I(g_{\boldsymbol{x}}) = \frac{1}{2} \log \frac{\prod V_{ii}}{\det \boldsymbol{V}}. \tag{37}$$

Let $\boldsymbol{A}$ be an invertible matrix such that $\boldsymbol{A}\,\boldsymbol{A}^{\mathsf{T}} = \boldsymbol{V}^{-1}$. After transform, the MI becomes $I(g_{\boldsymbol{Ax}}) = 0$. Hence, in order for the MI not to change, it is necessary to have $I(g_{\boldsymbol{x}}) = 0$, which is possible only in the particular cases where $\boldsymbol{V}$ is diagonal.

Note that for transforms $\boldsymbol{A}$ of the form $\boldsymbol{\Lambda P}$, where $\boldsymbol{\Lambda}$ is diagonal invertible and $\boldsymbol{P}$ is a permutation do not affect the MI.

To conclude, if $\boldsymbol{X} = \boldsymbol{Ax}$, it is in general not true that $\prod p_{x_i}(u_i) = \prod p_{X_j}(v_j)$. But we shall have scale invariance, as will be pointed out in proposition 9.

4.6.1 Negentropy

The differential entropy, or simply the *Entropy*, of a random variable admitting a density $p_{\boldsymbol{x}}(\boldsymbol{u})$ is defined as:

$$S(p_{\boldsymbol{x}}) \stackrel{\text{def}}{=} -\int p_{\boldsymbol{x}}(\boldsymbol{u}) \, \log p_{\boldsymbol{x}}(\boldsymbol{u}) \, d\boldsymbol{u}. \tag{38}$$

Example 12: Entropy of a Gaussian variate. in the real Gaussian case, the differential entropy takes the well-known form:

$$S(g_{\boldsymbol{x}}) = \frac{K}{2} \, (1 + \log 2\pi) + \frac{1}{2} \, \log \det \boldsymbol{V}$$

If entropy is defined with a logarithm in basis a (often $a = 2$), then it is easy to check out that $S(g_{\boldsymbol{x}}) = \frac{K}{2} \, \log_a 2\pi + \frac{K}{2 \log_e a} + \frac{1}{2} \, \log_a \det \boldsymbol{V}$

As an example, the MI is a difference beteen entropies:

$$I(p_{\boldsymbol{x}}) = \sum S(p_{x_i}) - S(p_{\boldsymbol{x}})$$

Entropy plays a special role in statistics. In fact, it is possible to show that there does not exist other functionals satisfying four basic very general axioms [126, page 27]. But it is convenient for our purposes to introduce the *negentropy*.

Definition 4 *Let $\boldsymbol{x}$ be a random variable admitting a density $p_{\boldsymbol{x}}(\boldsymbol{u})$. Denote $g_{\boldsymbol{x}}(\boldsymbol{u})$ the Gaussian density with same mean and covariance matrix. Then the negentropy associated with $p_{\boldsymbol{x}}$ is given by:*

$$J(p_{\boldsymbol{x}}) = \int p_{\boldsymbol{x}}(\boldsymbol{u}) \, \log \frac{p_{\boldsymbol{x}}(\boldsymbol{u})}{g_{\boldsymbol{x}}(\boldsymbol{u})} \, d\boldsymbol{u}. \tag{39}$$

It can be noticed that negentropy is nothing else but the divergence:

$$J(p_{\boldsymbol{x}}) = \mathrm{K}(p_{\boldsymbol{x}}, g_{\boldsymbol{x}}) \tag{40}$$

Hence, it measures some distance to normality. From the results we have seen so far, we also have the following:

Proposition 5 *The negentropy of a distribution $p_{\boldsymbol{x}}$ is always positive, and cancels if and only if $p_{\boldsymbol{x}}$ is Gaussian almost everywhere.*

More explicitly, we even have:

Proposition 6 *Negentropy is the difference between the two entropies:*

$$J(p_{\boldsymbol{x}}) = S(g_{\boldsymbol{x}}) - S(p_{\boldsymbol{x}}). \tag{41}$$

Proof. By definition of entropy, we have

$$\begin{aligned} S(g_{\boldsymbol{x}}) - S(p_{\boldsymbol{x}}) &= \int p_{\boldsymbol{x}}(\boldsymbol{u}) \log p_{\boldsymbol{x}}(\boldsymbol{u})\, d\boldsymbol{u} - \int p_{\boldsymbol{x}}(\boldsymbol{u}) \log g_{\boldsymbol{x}}(\boldsymbol{u})\, d\boldsymbol{u} \\ &\quad + \int p_{\boldsymbol{x}}(\boldsymbol{u}) \log g_{\boldsymbol{x}}(\boldsymbol{u})\, d\boldsymbol{u} - \int g_{\boldsymbol{x}}(\boldsymbol{u}) \log g_{\boldsymbol{x}}(\boldsymbol{u})\, d\boldsymbol{u}. \end{aligned}$$

Hence, it follows that $S(g_{\boldsymbol{x}}) - S(p_{\boldsymbol{x}}) = J(p_{\boldsymbol{x}}) + \int \log g_{\boldsymbol{x}}(\boldsymbol{u})\, [p_{\boldsymbol{x}}(\boldsymbol{u}) - g_{\boldsymbol{x}}(\boldsymbol{u})]\, d\boldsymbol{u}$. But the last term is null since by definition, $g_{\boldsymbol{x}}$ and $p_{\boldsymbol{x}}$ have same variance, and since $\log g_{\boldsymbol{x}}(\boldsymbol{u})$ is a polynomial of degree 2. □

Proposition 7 *Entropy and Negentropy are invariant by orthonormal changes of coordinates.*

Proof. Consider two random vectors, $\boldsymbol{x}$ and $\boldsymbol{y} = \boldsymbol{Q}\,\boldsymbol{x}$, where $\boldsymbol{Q}$ is invertible. Then entropy of $\boldsymbol{y}$ can be written as:

$$S(p_{\boldsymbol{x}}) = -\int p_y(\boldsymbol{Q}\boldsymbol{u})\, \log\left[|\det \boldsymbol{Q}|\, p_{\boldsymbol{y}}(\boldsymbol{Q}\boldsymbol{u})\right]\, |\det \boldsymbol{Q}|\, d\boldsymbol{u},$$

which yields the transformation rule for entropy:

$$S(p_{\boldsymbol{x}}) = S(p_{\boldsymbol{y}}) - \log|\det \boldsymbol{Q}|. \tag{42}$$

It is thus clear that entropy is invariant under any transform whose determinant is 1, and in particular orthogonal transforms. Negentropy is hence invariant on a set of transform at least as large, from proposition 6. □

The invariance set of Negentropy is actually larger than that of Entropy:

Proposition 8 *Negentropy is invariant by invertible basis change.*

Proof. One simply applies (42), which is valid for every invertible transform, to $p_{\boldsymbol{x}}$ et $g_{\boldsymbol{x}}$. By difference, $\log|\det \boldsymbol{Q}|$ disappears and $J(p_{\boldsymbol{x}}) = J(g_{\boldsymbol{x}})$. □

The following proposition states a useful decomposition of the MI, which is depicted in figure 4.

Proposition 9 *The Mutual Information is invariant by scale change, and can be decomposed as*

$$I(p_{\boldsymbol{x}}) = I(g_{\boldsymbol{x}}) + J(p_{\boldsymbol{x}}) - \sum_i J(p_{x_i}). \tag{43}$$

Proof. By definition of the MI:

$$I(p_{\boldsymbol{x}}) = \sum S(p_{x_i}) - S(p_{\boldsymbol{x}}). \tag{44}$$

Let $\boldsymbol{\Lambda}$ be an invertible diagonal matrix. Vector $\boldsymbol{\Lambda}\boldsymbol{x}$ has then entropy $I(p_{\boldsymbol{\Lambda}\boldsymbol{x}}) = \sum S(g_{x_i}) - \log \Lambda_{ii} - S(p_{\boldsymbol{x}}) + \log\det\boldsymbol{\Lambda}$, which proves that $I(p_{\boldsymbol{\Lambda}\boldsymbol{x}}) = I(p_{\boldsymbol{x}})$. Moreover, using property 6, relation (44) yields:

$$I(p_{\boldsymbol{x}}) = \sum S(g_{x_i}) - \sum J(p_{x_i}) - S(g_{\boldsymbol{x}}) + J(p_{\boldsymbol{x}}).$$

resorting another time to $I(g_{\boldsymbol{x}}) = \sum S(g_{x_i}) - S(g_{\boldsymbol{x}})$ eventually allows to conclude. □

In the above result, note that the Gaussian MI may be replaced by its expression (37).

This last property emphasizes the terms that may enter in a statistical dependence between components x_i. First of all $I(g_{\boldsymbol{x}})$, which is a contribution of order 2 (it involves only moments of order 2 or lower), and which may be canceled by a standardization as already seen earlier. We are left with only two terms in (43), which involve higher order moments. From proposition 7, the first negentropy is invariant by orthogonal transforms. The last term in (43) is consequently the only one that remains if only orthogonal transforms are allowed.

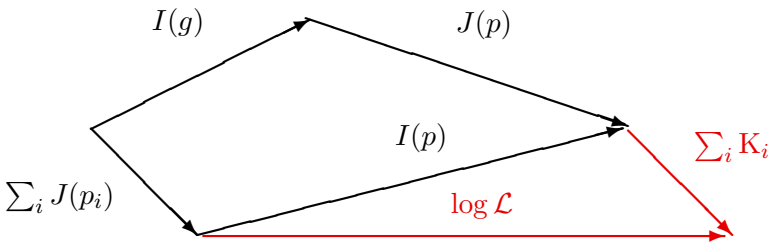


Figure 4: Decomposition of the Mutual Information and Log-Likelihood

However, the practical use of decomposition (43) is rather difficult, especially in large dimension (e.g. convolutive mixtures), even if some iterative algorithms have been devised [121]; the MI is nevertheless a powerful tool for separating Non Linear mixtures [136] [8]. Therefore, contrasts based on cumulants are often preferred in the context of linear mixtures.

For this purpose we shall see, in the following sections, how the Mutual Information can be approximated with the help of cumulants. We know already several things: (i) Blind problems cannot be solved under the sole assumption of statistical independence if observations are jointly Gaussian, (ii) cumulants measure some distance to Gaussianity, (iii) the gaussian distribution plays a central role in the MI decomposition we have suggested. We shall also see in the next section that (iv) the Gaussian distribution permits a computationally easier expansion.

4.6.2 Edgeworth expansion

Let x be a scalar random variable with second characteristic function $\Psi_x(u)$, supposed to be close to some other function $\Psi_o(u)$. By definition, $\Psi_x(u)$ generates the cumulants of x in its Taylor expansion about the origin:

$$\Psi_x(u) = \kappa_1 u + \frac{1}{2!}\kappa_2 u^2 + \frac{1}{3!}\kappa_3 u^3 + \dots, \tag{45}$$

where κ_r denotes the cumulant of order r, $\mathcal{C}_{x\,(r)}$. Denote λ_r the cumulant of order r in the expansion of $\Psi_o(u)$, and define the series $\eta_r = \kappa_r - \lambda_r$. The difference between characteristic functions can then be written as:

$$\Psi_x(u) - \Psi_o(u) = \sum_{r=1}^{\infty} \frac{1}{r!}\eta_r u^r. \tag{46}$$

Because the above difference is not necessarily a characteristic function, there may not exist a random variable having η_r as cumulants. But we may still denote μ_r the moments defined by the relation:

$$\exp[\sum_{r=1}^{\infty} \frac{1}{r!}\,\eta_r u^r] = \sum_{j=0}^{\infty} \frac{1}{k!}\,\mu_k u^k. \tag{47}$$

Now it is possible to expand the density $p_x(v)$ about $p_o(v)$

$$p_x(v) = p_o(v) \sum_{k=0}^{\infty} \frac{1}{k!} \mu_k h_k(v), \tag{48}$$

where functions $h_k(v)$ are defined as:

$$h_k(v) = \frac{(-1)^k}{p_o(v)} \frac{d^k p_o}{dv^k}(v). \tag{49}$$

Expansion (48) takes a simple form only for particular densities $p_o(v)$, especially when the functions $h_k(v)$ involved are polynomials.

The *Edgeworth expansion* of type A allows to approximate a density $p_x(v)$ when $p_o(v)$ is Gaussian. For consisency, we shall use the notation $p_o(v) = g_x(v)$ in that case, since we are indeed talking about the closed Gaussian distribution. For the sake of conciseness, and without restricting the generality, one may assume the density is zero mean and of unit variance. In that case, functions $h_k(v)$ are Hermite polynomials defined by the recursion:

$$h_0(v) = 1, \tag{50}$$
$$h_1(v) = v, \tag{51}$$
$$h_{k+1}(v) = v\, h_k(v) - \frac{d}{dv} h_k(v). \tag{52}$$

For instance, $h_2(v) = v^2 - 1$ and $h_3(v) = v^3 - 3v$.

The terms in the expansion can be ordered by increasing degrees, and this yields Gram-Charlier's expansion. But this gives no idea of how large the terms are and when they can be neglected. The other idea is to sort the successive terms of the expansion by their magnitude order under the assumptions of the Central Limit theorem (see page 19). The ordering is of no importance if all terms are kept, but becomes very important when a truncation is mandatory.

The central Limit theorem tells us that if x is a sum of m independent random variables with finite cumulants, then the rth order cumulant of x is of order $m^{1-r/2}$. This leads to the arrangement below:

Order							
$m^{-1/2}$	κ_3						
m^{-1}	κ_4	κ_3^2					
$m^{-3/2}$	κ_5	$\kappa_3\kappa_4$	κ_3^3				
m^{-2}	κ_6	$\kappa_3\kappa_5$	$\kappa_3^2\kappa_4$	κ_4^2	κ_3^4		
$m^{-5/2}$	κ_7	$\kappa_3\kappa_6$	$\kappa_3^2\kappa_5$	$\kappa_4^2\kappa_3$	κ_3^5	$\kappa_4\kappa_5$	$\kappa_3^3\kappa_4$

Hence, the Edgeworth expansion of a density $p_x(v)$ about $g_x(v)$ can be written as [87, formule 6.49]:

$$
\begin{aligned}
p_x(v)/g_x(v) &= 1 \\
&+\frac{1}{3!}\,\kappa_3 h_3(v) \\
&+\frac{1}{4!}\,\kappa_4 h_4(v) + \frac{10}{6!}\,\kappa_3^2 h_6(v) \\
&+\frac{1}{5!}\,\kappa_5 h_5(v) + \frac{35}{7!}\,\kappa_3\kappa_4 h_7(v) + \frac{280}{9!}\kappa_3^3 h_9(v) \\
&+\frac{1}{6!}\,\kappa_6 h_6(v) + \frac{56}{8!}\kappa_3\kappa_5 h_8(v) + \frac{35}{8!}\kappa_4^2 h_8(v) + \frac{2100}{10!}\kappa_3^2\kappa_4 h_{10}(v) \\
&\quad + \frac{15400}{12!}\kappa_3^4 h_{12}(v) \\
&+O(m^{-2}).
\end{aligned} \tag{53}
$$

Now it remains to use this expansion in order to approximate not only a distribution, but also a negentropy, and eventually a Mutual Information.

4.6.3 Approximation of Negentropy

We have seen with equation (37) that $I(g_x) = 0$ is and only if the covariance matrix is diagonal. For non Gaussian distributions, decorrelation at order 2 is not sufficient to ensure independence. On the other hand, negentropy is sufficient to characterize independence. However, as their probability distribution, negentropy of observations is in general unknown. It is proposed in this section to approximate it with the help of cumulants of increasing orders.

Let $p_x(u) = g(u)[1+f(u)]$, where $g(u)$ is the Gaussian distribution with zero mean and unit variance, as before, and $f(u)$ is given by the Edgeworth expansion. Take the following logarithm expansion $(1+f)\log(1+f) = f + f^2/2 - f^3/6 + f^4/12 + o(f^4)$ and plug it back into the Negentropy expression (39). Replacing $f(u)$ by its value, it is possible to obtain the expected approximation. In order to obtain the final result, it is necessary to use the following integral properties of Hermite polynomials:

$$\int g(v)\, h_p(v)\, h_q(v)\, dv = p!\,\delta_{pq}, \tag{54}$$

$$\int g(v)\, h_3^2(v)\, h_4(v)\, dv = 3!^3, \tag{55}$$

$$\int g(v)\, h_3^2(v)\, h_6(v)\, dv = 6!, \tag{56}$$

$$\int g(v)\, h_3^4(v)\, dv = 93\; 3!^2. \tag{57}$$

After some manipulations, one eventually gets (if x is indeed a random variable with zero mean and unit variance):

$$J(p_x) = \frac{1}{12}\kappa_3^2 + \frac{1}{48}\kappa_4^2 + \frac{7}{48}\kappa_3^4 - \frac{1}{8}\kappa_3^2\kappa_4 + o(m^{-2}). \tag{58}$$

4.6.4 The multivariate case

The previous expansion is valid in only the scalar case. But then how could one apply it for MIMO mixtures? The first obvious idea is to wonder whether this kind of result can be obtained in the

multivariate case. The answer is yes, up to a terrific increase in complexity. The result can be summarized as [35]:

$$J(p_{\boldsymbol{x}}) = \frac{1}{48}\left[\sum_{ijk}\kappa_{ijk}^2 + \sum_{ijk\ell}\kappa_{ijk\ell}^2 + 3\sum_{ijknqr}\kappa_{ijk}\kappa_{ijn}\kappa_{kqr}\kappa_{qrn} + 4\sum_{ijkmnr}\kappa_{ijk}\kappa_{imn}\kappa_{jmr}\kappa_{knr} - 6\sum_{ijk\ell m}\kappa_{ijk}\kappa_{i\ell m}\kappa_{jk\ell m}\right]$$

However, it turns out that this explicit result is not absolutely necessary in our framework when we handle *standardized variables*, and that the result in the scalar case can still be used in a smart way, as explained in the following section. On the other hand, if observations have not been standardized in a first stage, this is indeed that expression that should be utilized in (43).

5 Contrast functionals

The interest in using contrast optimization criteria appears in the presence of additive noise with unknown statistics, in which case it is impossible to resort to *Maximum Likelihood* (ML) approaches. In such cases, noise is merely seen as a nuisance, and can often be distinguished from useful source signals only by its comparatively smaller power.

However, as will be subsequently seen, the ML solution is linked to that of maximum contrast in the noiseless case, in particular if the Mutual information is utilized.

5.1 Assumptions

In order to blindly equalize convolutive models, the most widely used assumption is the *statistical independence* between successive symbols.

Hypothesis H1 *Sources $s_j[k]$ are all i.i.d. sequences.*

For MIMO models, the independence assumption between sources is often utilized:

Hypothesis H2 *Sources $s_j[k]$ are mutually statistically independent.*

These hypotheses can generally be deflated to less strong whiteness/independence properties, because moments of finite orders are used [42]. Let us stress that the case where sources are *linear processes* can also be treated in a similar manner as i.i.d. sources; Hypothesis H1 is thus not very restrictive. A particular case however raises problems, namely that of Gaussian sources. In that case, all the information is contained in moments up to order 2, which is not sufficient to establish identifiability. For this reason, it is necessary to resort to a third hypothesis, along with hypotheses H1 and H2:

Hypothesis H3 *At most one source is Gaussian.*

Again, this assumption can often be deflated to a weaker one. For instance, when the contrast optimization criterion is based on cumulants, then it is sufficient to consider the assumption

Hypothesis H4 *At most one source has a null marginal cumulant.*

On the other hand, there exist other frameworks in which hypotheses are different. For instance, if sources have different spectra, or if they are discrete, or non stationary, or cyclo-stationary, then they can be separated with the help of appropriate techniques.

The interest of exploiting the discrete character lies not only in a more accurate characterization of the desired output (than just non Gaussian or CM), but also in the fact that some other assumptions can be dropped. In this section, the sole assumption used is

Hypothesis H5 *The sources $s_j[n]$ belong to a known finite alphabet $\mathcal{A}$ characterized by the d distinct complex roots of a polynomial $Q(z) = 0$.*

For instance, sources can be correlated and non stationary. In fact, the criterion proposed above is entirely algebraic and deterministic, so that no statistical tool is required. The simplest case is $Q(z) = z^q - 1$, for which we have a PSK$-q$ constellation.

Next, for identifiability reasons, another hypothesis will be needed:

Hypothesis H6 *Sources $s_j[n]$ are sufficiently exciting*

By *sufficiently exciting*, it is meant that sufficiently many distinct states (among the d^P possible ones) of the $P-$uplet $\boldsymbol{s}$ are present in the source data matrix. For general alphabets, it is sufficient that all binary sequences are present, which means that the observation length should be at least of $\binom{d}{2}2^P$ symbols [41].

Hypothesis H7 *Each source $\boldsymbol{s}_i[n]$ is colored, and the set of the P spectra forms a family of P linearly independent functions.*

Hypothesis H8 *Sources are nonstationary, with linearly independent profiles.*

5.2 Trivial filters

The separating linear filter, $\check{\boldsymbol{F}}[z]$, if it exists, aims at delivering an output, $\boldsymbol{y}[n]$, which should satisfy as well as possible hypotheses H1 and H2. But it is clear that there exist some filters that do not affect them. These are called the trivial filters, and we can prove

Lemma 10 *Under hypotheses H1 to H3, trivial filters are of the form $\check{\boldsymbol{T}}[z] = \boldsymbol{P}\,\check{\boldsymbol{D}}[z]$, where $\boldsymbol{P}$ is a permutation, and $\check{\boldsymbol{D}}[z]$ a diagonal filter. In addition, because of the i.i.d. property of hypothesis H1, entries of $\check{\boldsymbol{D}}[z]$ must be of the form $\check{D}_{pp}[z] = \lambda_p\, z^{\delta_p}$, where δ_p is an integer.*

Consequently, it is hopeless to estimate the pair $(\check{\boldsymbol{H}}[z], \boldsymbol{s}[k])$. One should rather try to estimate one representative of the equivalence class of solutions. Once one solution is found, all the others can be generated by trivial filtering.

Example 13: Static model. Assume the model is MIMO static. Then $\boldsymbol{x}[n] = \boldsymbol{H}\,\boldsymbol{s}[n]$, where $\boldsymbol{x}[n]$ and $\boldsymbol{s}[n]$ are realizations of random variables. In that case, hypothesis H1 is not mandatory anymore. The estimation of the pair $(\boldsymbol{H}, \boldsymbol{s}[n])$ from the sole observations $\boldsymbol{x}[n]$ under hypotheses H2 and H3, has been introduced originally by Jutten and Comon [84] [32], and is now called *Independent Component Analysis* (ICA) [35] [118] [82] [99] [76] [39].

Consider now hypotheses H5 and H6 instead of hypotheses H1 to H3. In the PSK$-q$ case, we have the following

Lemma 11 *Under hypothesis H5 and H6 and with $Q(z) = z^q$, trivial filters are of the form $\boldsymbol{P}\,\check{\boldsymbol{D}}[z]$, where $\check{D}_{pp}[z]$ are rotations in the complex plane of an angle multiple of $2\pi/q$ combined with a pure delay, and $\boldsymbol{P}$ are permutations.*

For more general alphabets (called constellations in the framework of digital linear modulations), we need to define the following set

Definition 12 *Let $\mathcal{A}$ be a finite alphabet not reduced to $\{0\}$, defined by $Q(x) = 0$, where Q is a polynomial of degree d with $d > 1$ distinct roots, and let $\mathcal{G}$ be the subset of complex numbers γ, such that $\gamma\,\mathcal{A} \subset \mathcal{A}$.*

Note that, because $\mathcal{A}$ is finite, numbers of $\mathcal{G}$ are necessarily of unit modulus, and integer roots of unity (i.e., for every $\gamma \in \mathcal{G}$, there exists an integer $q(\gamma)$ such that $\gamma^{q(\gamma)} = 1$); this is true regardless of polynomial Q. In other words, $\gamma\,\mathcal{A} = \mathcal{A}$, $\forall \gamma \in \mathcal{G}$. Then, for every $\gamma \in \mathcal{G}$, the equation $\gamma\,\mathcal{A} = \mathcal{A}$ implies that $\gamma^{-1} \in \mathcal{G}$. Now, a similar statement as lemma 11 holds:

Lemma 13 *Trivial filters associated with hypotheses H5 and H6 are of the form $\boldsymbol{P}\,\check{\boldsymbol{D}}[z]$, where the entries of $\check{\boldsymbol{D}}[z]$ can be written as $D_{pp}[z] = \gamma_p\, z^n$, with $\gamma_p \in \mathcal{G}$ and $n \in \mathbb{Z}$.*

It can thus be seen that the size of the set of trivial filters (characterizing the inherent indeterminacy of the problem) depends on the nature of the alphabet.

5.3 Definition of contrasts

When noise is present in model (1), the estimation of inputs can be carried out according to a Maximum Likelihood (ML) or a Maximum A Posteriori (MAP) procedure if the noise has a known distribution. If this is not the case, noise must be considered as a nuisance. Contrast criteria are dedicated to this kind of situation.

Let $\mathcal{H}$ be a set of filters, and denote $\mathcal{H}\cdot\mathcal{S}$ the set of processes obtained by operation of filters of $\mathcal{H}$ on processes of $\mathcal{S}$. Denote $\mathcal{T}$ the subset of $\mathcal{H}$ of trivial filters, defined in lemma 10. An optimization criterion, $\Upsilon(\boldsymbol{H};\boldsymbol{x})$, is referred to as a contrast, defined on $\mathcal{H}\times\mathcal{H}\cdot\mathcal{S}$, if it satisfies the three properties below [37]:

P1 *Invariance:* The contrast should not change within the set of acceptable solutions, which means that
$\forall \boldsymbol{H} \in \mathcal{T}, \forall \boldsymbol{x} \in \mathcal{H}\cdot\mathcal{S}$, $\Upsilon(\boldsymbol{H};\boldsymbol{x}) = \Upsilon(\boldsymbol{I};\boldsymbol{x})$.

P2 *Domination:* If sources are already separated, any filter should decrease the contrast. In other words,
$\forall \boldsymbol{s} \in \mathcal{S}, \forall \boldsymbol{H} \in \mathcal{H}$, then $\Upsilon(\boldsymbol{H};\boldsymbol{s}) \leq \Upsilon(\boldsymbol{I};\boldsymbol{s})$.

P3 *Discrimination:* The maximum contrast should be reached only for filters linked to each other via trivial filters:
$\forall \boldsymbol{s} \in \mathcal{S}, \Upsilon(\boldsymbol{H};\boldsymbol{s}) = \Upsilon(\boldsymbol{I};\boldsymbol{s}) \Rightarrow \boldsymbol{H} \in \mathcal{T}$.

5.4 Kurtosis and skewness criteria

Most of the contrast criteria described in this section hold valid in the convolutive case. Therefore, they are described within this framework, without significant increase in complexity.

All the developments presented in the previous sections are valid for any random variable $\boldsymbol{x}$ with finite cumulants. We shall apply these results to variable $\boldsymbol{z}$ at the output of a separating filter $\boldsymbol{F}$, as depicted in figure 5. For this purpose, denote with symbol κ the source (standardized) kurtoses and with symbol γ the (standardized) cumulants of $\boldsymbol{z}$.

By inspection of decomposition (43) of the Mutual Information, it can be seen that only marginal negentropies $J(p_{x_i})$ need to be approximated when the joint negentropy $J(p_{\boldsymbol{x}})$ is invariant, which is

the case for orthonormal transforms. In such a case, it is sufficient to use approximation (58) for every component x_i, which yields [33] [35]:

$$I(p_{\boldsymbol{z}}) = J(p_{\boldsymbol{z}}) - \frac{1}{48}\sum_i 4\,\gamma_{iii}^2 + \gamma_{iiii}^2 + 7\,\gamma_{iii}^4 - 6\,\gamma_{iii}^2\,\gamma_{iiii} + o(m^{-2}). \tag{59}$$

where $\boldsymbol{x}$ is any standardized random variable (hence with $I(g_{\boldsymbol{x}}) = 0$) with bounded cumulants.

The first term, of order $O(m^{-1})$, is:

$$\Upsilon_{2,3} = \sum_{i=1}^{P} (\gamma_{iii})^2 \tag{60}$$

If 3rd order source cumulants are significantly larger than 4th order, then this quantity is a good approximation of the MI, up to the constant term $J(p_{\boldsymbol{x}})$. As a consequence, it is sufficient to maximize $\Upsilon_{2,3}$ in order to minimize the MI. On the other hand, if 3rd order cumulants are negligible, which happens for symmetrically distributed sources, then the first non zero term appears to be:

$$\Upsilon_{2,4} = \sum_{i=1}^{P} (\gamma_{iiii})^2 \tag{61}$$

which is a $O(m^{-2})$ term, in the sense of the Central Limit theorem.

One could also imagine to minimize the truncated expansion (59), which would be equivalent to minimizing the four terms in the sum, with respect to the orthonormal transform. But it turns out that we do not know whether this quantity is a contrast or not (and it is actually probably not).

In order to make sure that criteria (60) and (61) may be maximized to separate the sources, we must make sure they are indeed contrast functions. This is precisely the subject of the discussion below.

In the real case, that is, if sources, mixture, and noise are in the real field, then we have:

Proposition 14 *If at most one source has a null skewness, then the criterion*

$$\Upsilon_{2,3} = \sum_{p=1}^{P} (\gamma_{iii})^2$$

is a contrast.

Now in the real or complex cases, the following holds:

Proposition 15 *If at most one source has a null kurtosis, then the criterion*

$$\Upsilon_{2,4} = \sum_{p=1}^{P} (\gamma_{ii}^{ii})^2$$

is a contrast.

More generally:

Proposition 16 *If at most one source has a null standardized cumulant of order $r > 2$, the criterion below is a contrast for any real number $\alpha \geq 1$:*

$$\Upsilon_{\alpha,r} = \sum_{p=1}^{P} |\gamma_{(r)}|^{\alpha}$$

If all source kurtoses have the same sign, it is possible to drop the absolute value in contrast $\Upsilon_{1,4}$ (obtained for $\alpha = 1$) and get a contrast of simpler form:

Proposition 17 *If all source kurtoses have the same sign ε, and at most one is null, the following is a contrast:*

$$\Upsilon_{1,4} = \varepsilon \sum_{p=1}^{P} \gamma_{ii}^{ii}$$

This holds true for contrasts based on even order circular cumulants

Proof. We know that $\sum_i |\gamma_{ii}^{ii}|$ is a contrast, from Proposition 16. Let us show that this coincides with $\Upsilon_{1,4}$ defined in the above Proposition. Denote $\boldsymbol{G} = \boldsymbol{F}\,\boldsymbol{H}$ the global filter linking the outputs z_i of the separating filter $\boldsymbol{F}$ with the sources s_j. Then, by multi-linearity of cumulants, and because sources are independent, we have that $\gamma_{ii}^{ii} = \sum_j G_{ij} G_{ij} G_{ij}^* G_{ij}^* \kappa_{jj}^{jj}$, where κ_{jj}^{jj} are source kurtoses. Thus we have $\sum_i |\gamma_{ii}^{ii}| = \sum_i \left| \sum_j |G_{ij}|^4 \kappa_{jj}^{jj} \right|$. But all source kurtoses have the same sign ε, so that this also equals $\varepsilon \sum_i \sum_j |G_{ij}|^4 \kappa_{jj}^{jj}$, which is nothing else but $\Upsilon_{1,4}$. □

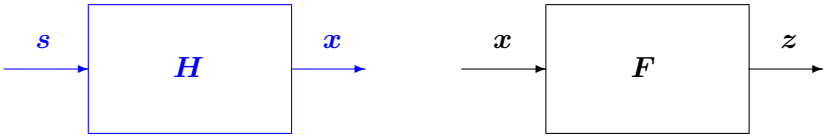


Figure 5: Contrast criteria aim at estimating a separating filter $\boldsymbol{F}$

Many other contrast criteria may be thought of, some of them being described in [114] [119] [41]. Let us mention only a simple way to build other extensions, before closing this subsection.

Proposition 18 *If $\Upsilon_k(\boldsymbol{y})$ are contrasts defined on $\mathcal{H}\cdot\mathcal{S}_k$, and $\{a_k\}$ are strictly positive numbers, then $\Upsilon(\boldsymbol{y}) \stackrel{\text{def}}{=} \sum_k a_k \Upsilon_k(\boldsymbol{y})$ is a contrast on $\mathcal{H}\cdot\bigcup_k \mathcal{S}_k$.*

See [41] for a proof. For instance, from Proposition 18, the criterion below is a contrast:

$$J_2^{(3+4)}(\boldsymbol{x}) = \sum_i 4\,|\gamma_{iii}|^2 + |\gamma_{iiii}|^2 \tag{62}$$

The interest in using proposition 18 is that the contrast built this way is discriminating over a wider set of sources. In fact, sources need to have *either* non-zero skewness or non-zero kurtosis.

To conclude, separating sources is equivalent to make the outputs of filter $\boldsymbol{F}$ "the least Gaussian possible", as already emphasized in 1980 by Donoho [57].

5.5 Joint Diagonalization contrasts

It may be interesting in some cases to build less efficient contrast criteria, with the goal of reducing the computational complexity of corresponding optimization numerical algorithms.

Let us explain this idea with the example of contrasts built on fourth order standardized cumulants. The best possible separation is achieved if all marginal output kurtoses are maximized, as accounted for in Proposition 15. A sub-optimal way of performing this maximization is to jointly diagonalize all the

3rd order slices. This idea gives birth to a recursion on the tensor order. Algorithms able to perform such maximizations are described in [96].

Note that, in the case when the exact diagonalization of the cumulant tensor is possible, diagonalizing all slices is equivalent to diagonalizing the tensor. This requires no noise, fewer sources than sensors, and a sufficiently many snapshots, hence the sub-optimality in actual problems.

The same limitations affect the so-called JADE algorithm, proposed in [23]. We shall see easily on the criterion itself why it is sub-optimal in the presence of noise. Denote Θ the Frobenius norm of the whole cumulant tensor of $\boldsymbol{z}$, $\gamma_{ij}^{k\ell}$. Then, by definition of $\Upsilon_{2,4}$:

$$\Theta - \Upsilon_{2,4} = \sum_{ijk\ell \neq iiii} |\gamma_{ij}^{k\ell}|^2$$

Instead of maximizing $\Upsilon_{2,4}$, one can thus minimize the above quantity consisting of all non diagonal entries.

If we decide to maximize some non diagonal entries, we obviously define a sub-optimal criterion. Suppose we define:

$$\Theta - \Upsilon_{Jade} = \sum_{ijk\ell \neq iik\ell} |\gamma_{ij}^{k\ell}|^2$$

Then, some entries are indeed maximized in

$$\Upsilon_{Jade} = \sum_{ik\ell} |\mathcal{C}_{i\ell}^{ik}|^2 \tag{63}$$

In addition, the four indices of the cumulant tensor, which played similar roles because of the (Hermitian or real) symmetries, now play different roles. The symmetry is consequently broken. But one advantage is that the above criterion can be written in compact form as:

$$\Upsilon_{Jade} = \sum_{p,s=1}^{P} ||\mathbf{diag}\{\boldsymbol{U}^{\mathsf{H}}\, \boldsymbol{M}(p,s)\, \boldsymbol{U}\}||^2 \tag{64}$$

if matrix $\boldsymbol{U}$ is unitary and where matrices $\boldsymbol{M}(r)$ are hermitian and defined as $M(r,s)_{pq} = \mathcal{C}_{ps}^{qr}$. In other words, the problem has been deflated to the Joint Approximate Diagonalization (JAD) of matrices.

5.6 Darmois's theorem

The following theorem has been independently discovered by Markinkievicz in 1940 and Dugué in 1951. It can be found in various textbooks of statistics, including [60] and [109].

Theorem 19 *The second characteristic function of a random variable cannot be a polynomial unless it is at most of degree 2.*

Another theorem, very useful for our concern, is due to Darmois (1953), and utilizes the above theorem:

Theorem 20 *Let two random variables be defined as linear combinations of independent random variables x_i:*

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

Then, if X_1 and X_2 are independent, those x_j for which $a_j b_j \neq 0$ are Gaussian.

The consequence of Darmois's theorem is given by the corollary below, and as pointed out in [35], it is important in the context of Independent Component Analysis.

Corollary 21 *If $\boldsymbol{z} = \boldsymbol{C}\,\boldsymbol{s}$, where s_i are non deterministic (their pdf is not reduced to a single mass) and independent, with at most one of them being Gaussian, then the following properties are equivalent:*

1. *Components z_i are pairwise independent*
2. *Components z_i are mutually independent*
3. *$\boldsymbol{C} = \boldsymbol{\Lambda}\,\boldsymbol{P}$, with $\boldsymbol{\Lambda}$ diagonal and $\boldsymbol{P}$ permutation*

This corollary allows to process outputs by pairs, as is proposed in many numerical algorithms operating after prewhitening.

However, it is important to insist that pairwise independence does not imply mutual independence, as explained by the two examples below.

Example 14: Simplest counter example of pairwise independence. We are given a bag containing 4 Bowls: 1 Red, 1 Yellow, 1 Green, 1 with the 3 colors. Denote $\{RB, YB, GB, RYGB\}$ this set of bowls. The bowls have equal drawing probabilities $P(RB) = P(YB) = P(GB) = P(RYGB) = 1/4$. Define the event "$R$" of drawing a bowl containing the Red color. We have $P(R) = P(RB) + P(RYGB) = 1/2$. By symmetry, we also have $P(G) = 1/2$ and $P(Y) = 1/2$, for Green and Yellow colors. Then $P(R \cap Y) = P(RYGB) = 1/4$ is equal to $P(R) * P(Y)$, which shows that events R and Y are pairwise independent. By symmetry, the two other pairs of color events are also pairwise independent. But $P(R \cap Y \cap G) = P(RYGB) = 1/4$ since it is the probability of drawing bowl $RYGB$. Yet, this is not equal to $P(R) * P(Y) * P(G) = 1/8$, which shows that the three color events are *not mutually independent.*

Example 15: Pairwise independence of virtual sources. The following property of BPSK variables has been used in [40] in order to extract 3 sources from a mixture received on 2 sensors. Consider 3 mutually independent BPSK sources, $x_i \in \{-1, 1\}$, $1 \leq i \leq 3$. Define $x_4 = x_1 x_2 x_3$. Then x_4 is also BPSK, *dependent on* x_i. One can verify that x_k are *pairwise independent*: $p(x_1 = a, x_4 = b) = p(x_4 = b \,|\, x_1 = a).p(x_1 = a) =$, $p(x_2\,x_3 = b/a).p(x_1 = a)$. But x_1 and $x_2\,x_3$ are BPSK, which yields that $p(x_2\,x_3 = b/a).p(x_1 = a) = \frac{1}{2} \cdot \frac{1}{2}$, showing pairwise independence. But x_k are obviously not mutually independent, $1 \leq k \leq 4$, because x_4 is built on the other x_i's. In particular, it can be shown that $\mathrm{Cum}\{x_1, x_2, x_3, x_4\} = 1 \neq 0$.

5.7 Maximum Likelihood

In the noiseless case, it is possible to link ML and MI criteria. As before, denote $p_{\boldsymbol{s}}(\cdot)$ the source joint pdf and $p_{\boldsymbol{x}}(\cdot)$ the joint pdf of observation. In the noiseless case, $\boldsymbol{x} = \boldsymbol{H}\,\boldsymbol{s}$ implies, by the rule of change of coordinates, that the likelihood of a measurement $\boldsymbol{x}_T$ writes:

$$p_{\boldsymbol{x}|\boldsymbol{H}}(\boldsymbol{x}_T|\boldsymbol{H}) = \frac{1}{|\det \boldsymbol{H}|}\, p_{\boldsymbol{s}}(\boldsymbol{H}^{-1}\boldsymbol{x}_T) \tag{65}$$

In the presence of an additive noise $\boldsymbol{v}$ of pdf $g(\cdot)$ independent of $\boldsymbol{s}$, the Likelihood takes the form:

$$p_{\boldsymbol{x},\boldsymbol{s}|\boldsymbol{H}}(\boldsymbol{x}_T, \boldsymbol{s}|\boldsymbol{H}) = g(\boldsymbol{x}_T - \boldsymbol{H}\,\boldsymbol{s}) \cdot p_{\boldsymbol{s}}(\boldsymbol{s})$$

And the *Joint MAP-ML* criterion to be used for a joint estimation of sources andmixture takes the form:

$$\begin{aligned}(\boldsymbol{s}_{MAP}, \boldsymbol{H}_{MV}) &= Arg \operatorname*{Max}_{\boldsymbol{s},\boldsymbol{H}} p(\boldsymbol{x}_T, \boldsymbol{s}|\boldsymbol{H}) \\ &= Arg \operatorname*{Max}_{\boldsymbol{s},\boldsymbol{H}} p(\boldsymbol{x}_T|\boldsymbol{s}, \boldsymbol{H})\, p_{\boldsymbol{s}}(\boldsymbol{s})\end{aligned}$$

In the noiseless case, the separating filter is nothing else but $\boldsymbol{F} = \boldsymbol{H}^{-1}$. Denote $p_{\boldsymbol{z}}(\cdot)$ the joint pdf of outputs of $\boldsymbol{F}$, that is, $\boldsymbol{z} = \boldsymbol{F}\boldsymbol{x}$. From (65), we have again that $p_{\boldsymbol{x}|\boldsymbol{H}}(\boldsymbol{u}|\boldsymbol{F}) = p_{\boldsymbol{x}|\boldsymbol{H}}(\boldsymbol{u}|\boldsymbol{H}) = |\det \boldsymbol{F}|\, p_{\boldsymbol{s}}(\boldsymbol{F}\boldsymbol{u})$. For an increasing number of independent observations, $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T\}$, the average log-likelihood can be seen to converge to

$$\mathcal{L}_T(\boldsymbol{H}) \stackrel{\text{def}}{=} \frac{1}{T} \log p(\boldsymbol{x}_1 \ldots \boldsymbol{x}_T|\boldsymbol{H}) \to \int p_{\boldsymbol{x}}(\boldsymbol{u})\, \log p_{\boldsymbol{x}|\boldsymbol{H}}(\boldsymbol{u}|\boldsymbol{H})\, d\boldsymbol{u}$$

Now add and subtract $\log p_{\boldsymbol{x}}(\boldsymbol{u})$ in order to have the required terms to build a *Kullback divergence* defined in (33), and get:

$$\mathcal{L}_\infty(\boldsymbol{H}) = -\mathrm{K}(p_{\boldsymbol{x}}, p_{\boldsymbol{x}|\boldsymbol{H}}) - S(p_{\boldsymbol{x}})$$

Then, because of the invariance of the Kullback divergence by invertible transform seen in property 3 page 23, this average likelihood can be eventually written as, up to the constant term $S(p_{\boldsymbol{x}})$ that does not depend on $\boldsymbol{F}$:

$$\Upsilon_{ML} \stackrel{\text{def}}{=} -\mathrm{K}(p_{\boldsymbol{z}}, p_{\boldsymbol{s}}) \tag{66}$$

This equation is interesting because, in contrast to the MI criterion, the ML criterion aims at minimizing the divergence between the pdf at the output $\boldsymbol{z}$ of the separating filter (here $\boldsymbol{F} = \boldsymbol{H}^{-1}$) and the source pdf.

Proposition 22 *The average log-Likelihood can be split into two parts thanks to the decomposition below:*

$$\mathrm{K}(p_{\boldsymbol{z}}, p_{\boldsymbol{s}}) = \mathrm{K}(p_{\boldsymbol{z}}, \prod_i p_{z_i}) + \sum_i \mathrm{K}(z_i, s_i) \tag{67}$$

Proof. By definition, $\mathrm{K}(p_{\boldsymbol{z}}, p_{\boldsymbol{s}}) = \int p_{\boldsymbol{z}} \log p_{\boldsymbol{z}} - \int p_{\boldsymbol{z}} \log p_{\boldsymbol{s}}$. Because $p_{\boldsymbol{s}} = \prod_i p_{s_i}$, the second term can be rewritten as $\sum_i \int p_{z_i} \log p_{s_i}$. Now add and subtract the term $\sum_i \int p_{\boldsymbol{z}} \log p_{z_i}$. This yields $\mathrm{K}(p_{\boldsymbol{z}}, p_{\boldsymbol{s}}) = \int p_{\boldsymbol{z}} \log \frac{p_{\boldsymbol{z}}}{\prod_i p_{z_i}} + \sum_i \int p_{\boldsymbol{z}} \log \frac{p_{z_i}}{p_{s_i}}$ □

In other words, as depicted in figure 4, the *log-Likelihood* contains the MI, plus a divergence between the output and source marginal pdf's. In the absence of knowledge of source pdf's, it is thus natural (and recommended) to use only the MI. On the other hand, if source distributions are known, the ML criterion will be more accurate.

6 Numerical algorithms for static MIMO mixtures

Because the BSS problem has become old since the nineties, there is a whole bunch of numerical algorithms available today. It is perhaps necessary to make a rough classification in a first stage.

One may distinguish between algorithms according to four features. First of all, one must make the distinction between Blind Identification algorithms, and Blind Source Separation/Extraction algorithms.

Next, as pointed out in section 2.2, algorithms may be fully adaptive, that is, recursive on time and calculating a new solution at each sample arrival. This type of algorithm has become less and

less attractive. One can also update the solution every time a block of N samples has arrived; the algorithm is adaptive block-wise. Conversely, on can process the data in a batch manner, that is when all the expected data have arrived. In practice, the difference between block-adaptive and purely batch is rather mild and somewhat arbitrary, since one can always use the estimate obtained with the previous block, at least as a prior guess.

The third feature of a numerical algorithm is the way sources are extracted: they can be extracted all together, in a joint manner (and often in a symmetric manner as well). Nevertheless, many techniques involve sweeping all possible pairs of outputs. In this case, the joint extraction algorithm calls a pair-wise processing subroutine. The other possibility is to extract sources one by one, which is often referred to as *Deflation*.

Lastly, it is important to check out whether the algorithms need a prior standardization (whitening) or not. For instance, BI of Under-Determined mixtures does not resort to data standardization.

The goal is not to give here an exhaustive list of all available algorithms. Instead, only a few examples are described, as an illustration.

6.1 Closed-form separation with $\Upsilon_{1,4}$ for 2×2 mixtures

The simplest examples leading to closed-form solutions even in the presence of noise are found with 2 sources with same kurtosis sign, when maximizing contrast $\Upsilon_{1,4}$ defined in Proposition 17 page 33, after prior standardization. It is convenient to address the real and complex cases separately.

6.1.1 2×2 real mixtures

Since standardization has been performed beforehand, one looks for an estimate $\boldsymbol{z}$ of sources, expressed in the form of an orthonormal transform of standardized data:

$$\boldsymbol{z} = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix} \tilde{\boldsymbol{x}} \tag{68}$$

Denote with symbol κ the output kurtoses, and with γ those of standardized data. Also assume the short notation $\kappa_i = \kappa_{iiii}$ and $\gamma_i = \gamma_{iiii}$. The multi-linearity of cumulants yields the following input-output relations:

$$\begin{aligned} \kappa_1 &= \gamma_1 \cos^4\alpha + 4\gamma_{1112}\cos^3\alpha\sin\alpha + 6\gamma_{1122}\cos^2\alpha\sin^2\alpha \\ &+ 4\gamma_{1222}\cos\alpha\sin^3\alpha + \gamma_2\sin^4\alpha \\ \kappa_2 &= \gamma_1 \sin^4\alpha - 4\gamma_{1112}\cos\alpha\sin^3\alpha + 6\gamma_{1122}\cos^2\alpha\sin^2\alpha \\ &- 4\gamma_{1222}\cos^3\alpha\sin\alpha + \gamma_2\cos^4\alpha \end{aligned}$$

Then it is possible to prove, after some basic manipulations, that:

$$\begin{aligned} \varepsilon\Upsilon_{1,4} &= \kappa_1 + \kappa_2 = \\ & [\cos 2\alpha \ \sin 2\alpha] \begin{pmatrix} \gamma_1+\gamma_2 & \gamma_{1112}-\gamma_{1222} \\ \gamma_{1112}-\gamma_{1222} & \frac{\gamma_1+\gamma_2}{2}+3\gamma_{1122} \end{pmatrix} \begin{bmatrix} \cos 2\alpha \\ \sin 2\alpha \end{bmatrix} \end{aligned}$$

The maximum of the contrast is thus obtained when the unit vector is the dominant eigenvector of the above matrix. It can be computed via EVD for instance. Then angle α can be calculated, up to indeterminacies entering the inherent scale-permutation indeterminacies of the BSS problem.

6.1.2 2×2 complex mixtures

Let us now turn to complex mixtures; vector $\boldsymbol{z}$ is now a unitary combination of standardized observations:

$$\boldsymbol{z} = \begin{pmatrix} \cos\alpha & \sin\alpha\, e^{\jmath\varphi} \\ -\sin\alpha\, e^{-\jmath\varphi} & \cos\alpha \end{pmatrix} \tilde{\boldsymbol{x}} \tag{69}$$

Again define κ_i the output (circular) kurtoses, $\mathcal{C}_{\boldsymbol{z}}{}_{ii}^{ii}$, $\gamma_i = \mathcal{C}_{\tilde{\boldsymbol{x}}}{}_{ii}^{ii}$ and $\gamma_{ij}^{k\ell} = \mathcal{C}_{\tilde{\boldsymbol{x}}}{}_{ij}^{k\ell}$. Then contrast $\Upsilon_{1,4}$ is again proved to be a quadratic form in some unit vector depending on α [38]:

$$\varepsilon\Upsilon_{1,4} \;=\; \kappa_1 + \kappa_2 = \boldsymbol{u}^{\mathsf{T}}\, \boldsymbol{B}\, \boldsymbol{u}$$

with

$$\boldsymbol{u}^{\mathsf{T}} = [\cos 2\alpha \quad \sin 2\alpha\cos\varphi \quad \sin 2\alpha \sin\varphi]$$

and

$$\begin{aligned} \boldsymbol{B} &= \begin{pmatrix} \gamma_1+\gamma_2 & \Re\{\delta\} & -\Im\{\delta\} \\ \Re\{\delta\} & 2\gamma_{12}^{12} + \Re\{\gamma_{22}^{11}\} & \Im\{\gamma_{22}^{11}\} \\ -\Im\{\delta\} & \Im\{\gamma_{22}^{11}\} & 2\gamma_{12}^{12} - \Re\{\gamma_{22}^{11}\} \end{pmatrix}; \\ \delta &= \gamma_{12}^{11} - \gamma_{22}^{12} \end{aligned}$$

The contrast maximum is thus obtainable in closed-form, since the EVD of a real 3×3 matrix can be computed entirely algebraically (it is not iterative). Angles α and ϕ are then obtained up to an inherent indeterminacy, yielding eventually the unitary matrix.

The algorithm based on this algebraic solution will be referred to as *CoM1*. Other more complicated algebraic solutions have been proposed before, and names such as CM or CoM have been used [47] [36]. The the terminology CoM*d* for algebraic solutions maximizing $\Upsilon_{d,4}$ makes sense, and is assumed subsequently.

6.2 Jacobi sweeping

We have already described in section 3.2 the standard cyclic Jacobi sweeping for symmetric (or Hermitian) matrices. The idea is similar for sweeping all the pairs of a symmetric tensor (the number of pairs of indices is the same, since the sweeping is performed in a symmetric manner). It has been originally proposed in [31] for the BSS problem. See [36] for convergence issues.

To make it simple, let us describe one sweep of a $3 \times 3 \times 3$ real symmetric tensor

$$\begin{pmatrix} X & x & x \\ x & x & x \\ x & x & . \end{pmatrix} \begin{pmatrix} x & x & x \\ x & X & x \\ x & x & . \end{pmatrix} \begin{pmatrix} x & x & x \\ x & x & x \\ x & x & . \end{pmatrix} \rightarrow \begin{pmatrix} X & x & x \\ x & . & x \\ x & x & x \end{pmatrix} \begin{pmatrix} x & x & x \\ x & . & x \\ x & x & x \end{pmatrix} \begin{pmatrix} x & x & x \\ x & . & x \\ x & x & X \end{pmatrix} \rightarrow \begin{pmatrix} . & x & x \\ x & x & x \\ x & x & x \end{pmatrix} \begin{pmatrix} . & x & x \\ x & X & x \\ x & x & x \end{pmatrix} \begin{pmatrix} . & x & x \\ x & x & x \\ x & x & X \end{pmatrix}$$

Symbols X indicate entries that are maximized, x the entris that are minimized, and $\cdot$ entries remaining unchanged by the last Givens rotation applied on the three tensor ways. As for matrices, there are 3 pairs to process in a sweep over a 3-dimensional symmetric tensor.

There is an obvious interpretation to make at this point. In order to fix the ideas, take the case of $\Upsilon_{2,4} = \sum_p |\kappa_{ii}^{ii}|^2$, withour restricting the generality. Under the action of unitary matrices, it can be easily shown that the quantity

$$\Theta = \sum_{abcd} |\kappa_{ab}^{cd}|^2$$

remains constant. It is in fact the Frobenius norm of the tensor κ. In other words, maximizing $\Upsilon_{2,4}$ is equivalent to minimizing the sum of squares of non diagonal terms. Thus, the procedure is nothing else but a *Tensor Diagonalization* attempt.

However, compared to matrices, a major difference is that none of the entries is guaranteed to be canceled out, contrary to matrices. In fact, not every tensor can be diagonalized by linear invertible change of coordinates; only rank deficient tensor can. This will be deepen in a subsequent section.

Similarly, the *JADE* algorithm, maximizing criterion (63), is in general not able to diagonalize exactly jointly a set of (more than two) matrices. The diagonalization is thus in general also approximate.

6.3 Deflation

The deflation approach of BSS consists of searching for an extracting vector, $\boldsymbol{f}$, so that its (single) output

$$z \stackrel{\text{def}}{=} \boldsymbol{f}^\mathsf{T}\boldsymbol{x}$$

maximizes some contrast function. This allows to extract one source. In a second stage, the contribution of this source can be subtracted in the observation by linear regression. The number of sources present has been reduced by one, and the process can keep going until all sources have been extracted.

Consider for instance the kurtosis criterion below, valid if data, mixture, and sources are real (the problem in the complex field is very similar, but a little complicated by notations):

$$\Upsilon(\boldsymbol{f}) = \frac{\mathrm{E}\{(\boldsymbol{f}^\mathsf{T}\boldsymbol{x})^4\}}{\mathrm{E}\{(\boldsymbol{f}^\mathsf{T}\boldsymbol{x})^2\}^2} - 3$$

where we have assumed that $\boldsymbol{s}$, $\boldsymbol{x}$ and z are zero-mean random variables. Because the BSS problem is known to be identifiable only up to a scale factor, it is not a surprise to see that the kurtosis criterion is insensitive to the norm of $\boldsymbol{f}$. It is thus necessary to impose a constraint on vector $\boldsymbol{f}$; this can be either $||\boldsymbol{f}||^2 = 1$ or $\mathrm{E}\{(\boldsymbol{f}^\mathsf{T}\boldsymbol{x})^2\} = 1$.

Under the latter constraint, stationary values of Υ are obtained when the objective and constraint differentials are collinear:

$$\mathrm{E}\{\boldsymbol{x}\,(\boldsymbol{x}^\mathsf{T}\boldsymbol{f})^3\} = \lambda\,\mathrm{E}\{\boldsymbol{x}\,(\boldsymbol{x}^\mathsf{T}\boldsymbol{f})\}$$

where the Lagrange multiplier can be shown to be $\lambda = \mathrm{E}\{z^4\}/\mathrm{E}\{z^2\}$.

If Deflation is run after prior standardization, which is not necessary, then the two above mentioned constraints are equivalent, and the kurtosis criterion reduces to a moment criterion, up to a constant:

$$\Upsilon(\boldsymbol{f}) = \mathrm{E}\{(\boldsymbol{f}^\mathsf{T}\boldsymbol{x})^4\}$$

The condition of collinearity of differentials then becomes:

$$\mathrm{E}\{x_i\,(\boldsymbol{f}^\mathsf{T}\boldsymbol{x})^3\} = \lambda\,f_i,\ \forall i \tag{70}$$

for some Lagrangian multiplier λ to be determined. This would result in a fixed-point equation if λ was known. Actually, it must be determined so as to satisfy the constraint, and thus unfortunately depends

again on $\boldsymbol{x}$ and $\boldsymbol{f}$. In [82] [80], λ is arbitrarily set to a deterministic fixed value, which allows to spare computations. For this reason, as pointed out in [81], the *FastICA* algorithm should be viewed rather as a Newton approximation than a fixed point algorithm.

Beside this approximation of FastICA, there exist other deflation algorithms. One of the very first is that of Delfosse and Loubaton [52], where the filter $\boldsymbol{f}$ is parametrized by angles of plane rotations (because a unit norm vector is nothing else but a column of an orthonormal matrix). This algorithm is thus executed after standardization. Simple gradient ascent can be also implemented, like the fixed step gradient as in [145]. There exist other solutions making fewer approximations, like the Robust ICA suggested in [41], where the optimal step size is calculated at every iteration.

7 Multi-way Factor Analysis

The usefulness to use Multi-way factor analysis may be argued in different manners. Taking into account what has been said so far, the easiest argument to put forward is the case of *Under-Determined Mixtures (UDM)*.

These mixtures are such that the number of sources, P, strictly exceeds the number of sensors, K. The consequence is that this mixture obviously does not admit a linear inverse. Thus, there is no possibility to define contrast criteria as we did so far, or the definition would need to be strongly revisited.

Another argument is more difficult to explain at this stage. Let's still introduce the necessary material, and we shall go back to this later on. In Antenna Array Processing, data are often recorded for different values of time and space. That's why they may be arranged in a matrix (array of order 2). Suppose the data are measured according to more than two indices, say space, time, and frequency for instance. Then a natural way to store them would be in a 3-way array. This arrangement is meaningful if frequency slices are not all proportional to each other. If this was the case, one could keep only one matrix slice without loosing information. One often says that storing in a different way is recommended when frequency *diversity* is present.

This discussion suggests that there exist two possible approaches:

- Using the Cumulant tensor of the 2-way data if no additional diversity is present
- Decomposing directly the r-way data array if sufficient diversity is present

In both cases, it will be necessary to decompose a tensor of order larger than 2 into elementary components. This may be viewed as one of the possible extensions of SVD to higher order arrays. The main difference between the two approaches is that the Cumulant array is a *square tensor*, that is, all its dimensions are equal, and that it also enjoys symmetries. On the other hand, the data array may have different dimensions.

7.1 Canonical Decomposition

We have already seen before that cumulants enjoy the multi-linearity property. This allows them to deserve the name of *tensors*. If data follow the noiseless linear model:

$$\boldsymbol{x} = \boldsymbol{H}\,\boldsymbol{s}$$

where sources s_i are statistically independent, then the cumulant tensor of observations x_j can be decomposed into (taking the example of the 3rd order):

$$\mathcal{C}_{\boldsymbol{x}\,ijk} = \sum_p H_{ip}\,H_{jp}\,H_{kp}\,\mathcal{C}_{\boldsymbol{s}\,ppp}$$

More generally, given any symmetric tensor $\boldsymbol{T}$, say of order 3, then its Canonical Tensor Decomposition (CanD) is defined as:

$$\boldsymbol{T} = \sum_{p=1}^{rank(\boldsymbol{T})} \kappa_p \, \boldsymbol{h}(p) \circ \boldsymbol{h}(p) \circ \boldsymbol{h}(p) \tag{71}$$

where $rank(\boldsymbol{T})$ denotes the smallest integer for which the decomposition is exact. Therefore, every tensor in the right hand side can be called a *rank one tensor*.

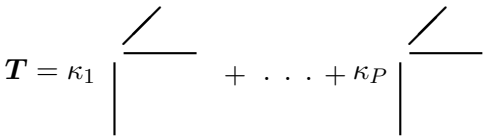


Figure 6: Canonical Decomposition of a 3rd order tensor

Now for a tensor of general form, T_ijk, where $1 \leq i \leq I$, $1 \leq j \leq J$, $1 \leq k \leq K$, the CanD may be expressed similarly as:

$$\boldsymbol{T} = \sum_{p=1}^{rank(\boldsymbol{T})} \boldsymbol{a}(p) \circ \boldsymbol{b}(p) \circ \boldsymbol{c}(p) \tag{72}$$

where the scalar weighting factor κ_p has been pulled into one of the loading vector terms. With the notation introduced in section 3.3, this can be written in compact form:

$$\boldsymbol{T} = \boldsymbol{A} \underset{2}{\bullet} \boldsymbol{B} \underset{2}{\bullet} \boldsymbol{C}$$

Matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ are the loading matrices, of size $I \times P$, $J \times P$, and $K \times P$, respectively, if $P = rank(\boldsymbol{T})$.

The concept of *tensor rank* has been apparently well defined. But there are several striking differences with matrix rank that need to be stressed. First of all, the rank of a real tensor may not be the same if computed in the real or complex fields, as shows the example below.

Example 16: Complex rank:.

$$\boldsymbol{T}(:,:,1) = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \boldsymbol{T}(:,:,2) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

If decomposed in $\mathbb{R}$, it is of rank 3:

$$\boldsymbol{T} = \frac{1}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix}^{\circ 3} + \frac{1}{2} \begin{pmatrix} 1 \\ -1 \end{pmatrix}^{\circ 3} - 2 \begin{pmatrix} 1 \\ 0 \end{pmatrix}^{\circ 3}$$

whereas it admits a CanD of rank 2 in $\mathbb{C}$:

$$\boldsymbol{T} = \frac{\jmath}{2} \begin{pmatrix} -\jmath \\ 1 \end{pmatrix}^{\circ 3} - \frac{\jmath}{2} \begin{pmatrix} \jmath \\ 1 \end{pmatrix}^{\circ 3}$$

This is thus extremely important to indicate in which field the rank is computed. Of course, for any real tensor $\boldsymbol{T}$, it holds true that:

$$rank_{\mathbb{C}}(\boldsymbol{T}) \leq rank_{\mathbb{R}}(\boldsymbol{T})$$

Similarly, we do not know whether the rank of a symmetric tensor is always the same if the rank-one tensors involved in the CanD are imposed to be symmetric. Therefore, this constraint will be assumed in the sequel.

Lastly, contrary to matrices, it is difficult to compute the rank of a given tensor, since we have not at our disposal a numerical algorithm able to compute the CanD (or even the rank only) in the general case. The algorithms available today are usable only under constraints concerning either the dimensions or the rank itself.

7.2 Decomposition of the Cumulant Tensor

We restrict our attention in this section to symmetric tensors, that is, tensors for which $T_{ijk} = T_{\sigma(i,j,k)}$ for any permutation σ among indices.

Let us define the *typical rank* ω of symmetric tensors of order d and dimension K: this is the rank that such tensors will have with probability 1 if their entries were drawn independently according to a continuous distribution on the real line.

Then it can be shown [48] that the typical rank computed in $\mathbb{C}$ of real symmetric tensors takes the values given in the table 2 below.

ω	K	**2**	**3**	**4**	**5**	**6**	**7**	**8**
d	**3**	2	4	5	8	10	12	15
	4	3	6	10	15	22	30	42

Table 2: Generic rank ω of symmetric tensors as a function of the dimension K and the order d

It can be read for instance in the table that a $3 \times 3 \times 3$ symmetric tensor has generically a rank 4 in $\mathbb{C}$. However, its CanD is not unique, as reported in table 3.

D	K	**2**	**3**	**4**	**5**	**6**	**7**	**8**
d	**3**	0	2	0	5	4	0	0
	4	1	3	5	5	6	0	6

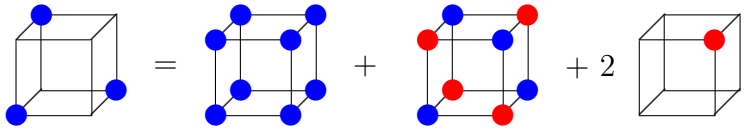
Table 3: Generic dimension D of the manifold of solutions

At this point, one can stress additional striking differences with matrices. First, the typical rank of a symmetric tensor is generally larger than its dimension. Next, the CanD of a typical tensor is *essentially unique* (i.e. unique up to scale and permutation) only for some values of order and dimension; see [48] for values for other orders or dimensions. Last, there exist tensors having a rank larger than the typical rank. In order to show this, just take the example below [39]:

Example 17: Tensor of order 3, dimension 2, but rank 3. Define the $2 \times 2 \times 2$ symmetric tensor, null everywhere except for the entries $T_{112} = T_{121} = T_{211} = 1$. This tensor admits the decomposition:

$$\boldsymbol{T} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}^{\circ 3} + \begin{pmatrix} -1 \\ 1 \end{pmatrix}^{\circ 3} - 2 \begin{pmatrix} 0 \\ 1 \end{pmatrix}^{\circ 3}$$

which is an explicit irreducible CAND. This decomposition is depicted in figure below.



blue bullets = 1, red bullets = −1.

In dimension 2, CanD entirely computable thanks to Sylvester's theorem on polynomials, in $\mathbb{R}$ or in $\mathbb{C}$. But the calculation of a CanD is very hard in higher dimensions.

Example 18: Tensor of order 3, dimension 3, but rank 5. Similarly, define the tensor of order 3 and dimension 3 by $T_{ijk} = 0$ everywhere except for the entries for which $i+j+k = 7$, for which the value is 1. It can be shown that this tensor has rank 5. Note that the typical rank is 4 in that case.

Does BI yield the sources?

In general, the Blind Identification of the UDM does not yield the sources. The problem is just changed into another one, once the mixture is estimated. In fact, the problem then consists of estimating the inputs of a known linear system whose outputs are observed. But there are cases where it is possible to estimate the sources perfectly, and in particular when they have a discrete distribution. It is even sometimes possible to extract directly the sources without prior BI, as shows the example commented now.

Let us go back to example 5.6 page 35. We defined a system with 3 inputs and 2 outputs. Inputs s_i take their values in $\{-1, 1\}$, and are mutually statistically independent. In order to be able to build a linear inverse, it is necessary to increase the size of the observation space, in a non linear manner. This is made by building the virtual observations [40]: $\boldsymbol{z} = [x_1^3, x_1^2x_2, x_1x_2^2, x_2^3]^\intercal$. Then the new system to invert is 4×6, and thus not Under-Determined anymore:

$$\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{z} \end{pmatrix} = \begin{bmatrix} \boldsymbol{H} & \boldsymbol{0} \\ \boldsymbol{B} & \end{bmatrix} \cdot \begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_1s_2s_3 \end{pmatrix}$$

with one virtual source $s_4 \stackrel{\text{def}}{=} s_1s_2s_3$. The BSS problem is solvable because source s_4 is pairwise independent from the others.

7.3 Decomposition of the Data Tensor

In an increasing number of actual problems, it becomes very useful to be able te decompose a non symmetric tensor into a sum of rank-one terms. The tensor to be decomposed can be the data recorded themselves [132], or sample statistics derived from them [49]. The direct decomposition of the Data tensor finds many applications, mainly in *Factor Analysis* [132], but also more recently in Sensor Array Processing [128] [27].

In practice, instead of trying to fit exactly the data tensor with a CanD, it is preferred to approximate the data tensor by another of lower rank:

$$\boldsymbol{T} = \sum_{p=1}^{\omega} \kappa_p \, \boldsymbol{a}(p) \circ \boldsymbol{b}(p) \circ \boldsymbol{c}(p) + \boldsymbol{E} \tag{73}$$

This is known as the *CP decomposition*, and was introduced independently by Harshman [75] and Carroll [24]. See [88] [132] [15] [46] for a general introduction to CP decompositions.

The problem is, given $\boldsymbol{T}$, to compute an estimate of loading matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$. One may restrict to the third order for the moment, in order to ease understanding. The error $\Psi \stackrel{\text{def}}{=} ||\boldsymbol{E}||^2$ is an homogeneous polynomial of degree 6 in many variables. Stationary points are thus solutions of a system of polynomials of degree 5 in many variables. There are algebraic techniques dedicated to thus kind of problem, such as the Groebner bases, but they are practicable only for very limited values of the degree and number of variables.

The CP decomposition can be computed by *Alternating Least Squares* (ALS). In fact, suppose $\boldsymbol{B}$ and $\boldsymbol{C}$ are fixed. Then error Ψ is quadratic in each $\boldsymbol{a}(p)$, and can then be minimized algebraically (calling for the SVD for instance). The idea is just to run a relaxation over these 3 modes, and estimate in turn $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$.

Other useful writings of the criterion to be minimized include:

$$\Psi = \sum_{k=1}^{K_3} ||\boldsymbol{T}(:,:,k) - \boldsymbol{A}\, Diag(\boldsymbol{C}(k,:))\, \boldsymbol{B}^{\mathsf{T}}||^2$$

or

$$\Psi = ||\boldsymbol{T}^{(1)} - \boldsymbol{A}(\boldsymbol{C} \odot \boldsymbol{B})^{\mathsf{T}}||^2 \tag{74}$$

This ALS algorithm is of course iterative, and unfortunately known to converge extremely slowly. In addition, there is no proof of global convergence (the minimum reached might be local is some cases). There are other algorithms that have been proposed, but none of them is entirely satisfactory; see [46] [14] references therein.

Kruskal rank and essential uniqueness

Given a matrix $\boldsymbol{A}$, we have $rank(\boldsymbol{A}) = r$ iff there is *at least one* subset of r lin. independent columns, and if this *fails for any subset* of $r+1$ columns. In contrast, one defines the *Kruskal rank* of a matrix, $k(\boldsymbol{A})$: $r_K(\boldsymbol{A}) = k$ iff *every* subset of k columns is lin. independent, and this *fails for at least one subset* of $k+1$ columns.

Property: An immediate consequence is that $k(\boldsymbol{A}) \leq rank(\boldsymbol{A}) \leq dim(\boldsymbol{A})$.

According to [89] [127] [13], the uniqueness of the CP analysis can be ensured if some sufficient conditions are satisfied. The sufficient condition proposed by Kruskal in 1977 is

$$k(\boldsymbol{A}) + k(\boldsymbol{B}) + k(\boldsymbol{C}) \geq 2\,\omega + 2$$

where $k(\boldsymbol{A})$ denotes *Kruskal's rank* of $\boldsymbol{A}$.

This condition attempts to express the need for *diversity*: matrix slices must be "sufficiently different". This Uniqueness sufficient condition can be extended to order d as: $2\omega \leq dK - d + 1$. For instance in the symmetric case, one would need at least that $2\omega \leq 3K - 2$, which is almost never satisfied by the typical rank.

On the other hand, a sufficient condition that allows almost surely a unique CP decomposition is that the rank is strictly smaller than its "expected value" [46].

7.4 Unexpected topological properties

In order to approximate in the Least Squares (LS) sense an object of an Euclidian vector space by another belonging to a subset $\mathcal{S}$, it is required that $\mathcal{S}$ be closed (in the sense of the topology induced by the norm). But is this condition indeed verified for the CP analysis?

It turns out that the answer is no, unfortunately. More precisely, denote $\mathcal{Y}_r$ the subset of tensors of rank at most r. Then $\mathcal{Y}_r$ is closed only for $r = 1$. In other words, it is theoretically not possible to find a LS rank-r approximate of a given tensor, unless $r = 1$. It is even possible to find a series of rank r tensors converging towards a tensor of rank $r + 1$ [43]. This result is very surprising, when we base our intuition on the algebra of matrices!

This is currently the subject of researches, which have not yet allowed to sketch an solid theory on the subject. This is the reason why it will not be elaborated on this any further, even if the CP has been widely used for over thirty years...

8 Concluding remarks

Contrary to ICA, which is now a subject well defined that has been widely studied, Tensor Decompositions are just beginning to raise increasing interest. Therefore, these notes have only tackled this strange subject, and have rather raised problems than solved. However, even if little is already known about identifiability conditions of the CP decomposition, and necessary conditions for uniqueness, one must recognize that CP is widely used with success, with the help of various suboptimal numeriacl algorithms, including ALS.

Multi-linear vs Linear Blind Model fitting

As we already stressed, CanD is generally necessary when the number of sources exceeds the number of sensors, even if we gave an exception with BPSK sources (and there are certainly other exceptions). If there is enough diversity among loading vectors, a multi-way data tensor can be built, which allows to use the CP decomposition (cf. figure 7). Otherwise, the data may be stored in matrix form without loosing information, and need to be treated with the ICA tools (cf. figure ica-fig).

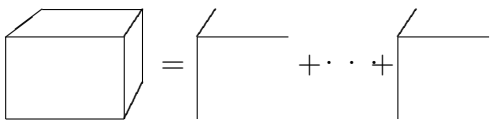


Figure 7: Canonical decomposition (CanD) of a Data -or Cumulant- tensor.

Some Unaddressed Problems

This tutorial was very incomplete, and it is important to stress that. A number of subjects that have been left aside can be listed, and include the following.

- There are other natural possible extensions of SVD to tensors. One of them is the so-called called *HOSVD* [93] [95]. In the HOSVD, the decomposition constructed with matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ does not involve a diagonal tensor but a *Core tensor* which has all-orthogonal properties. It is diagonal only in the matrix case. This is one of the instances of the *Tucker3* modeling [102] [143] [132] [46].

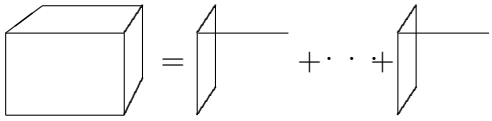


Figure 8: BI decomposes a data tensor as if they were stored in matrix format, ignoring that there are more than two ways. There is nothing else that can be done if diversity lacks. If the rank does not exceeds the dimension, a contrast-based approach can be assumed, and ICA may be run. Otherwise, BI of the UDM must be performed, via the construction of a Cumulant tensor whose CanD will be computed.

- As far as tensor numerical algorithms are concerned, there are some alternatives that have been proposed. One of them is the recursion on the tensor order, with the Simultaneous lower-order Tensor Diagonalization (STD) [96] [44, Ch.3].

- Nothing has been said on Performance Indices. For invertible mixtures, the Signal to Interference plus Noise Ratio (SINR) can be used, keeping in mind that there is a permutation ambiguity, which makes things more complicated. For UDM or CanD in the presence of more sources than sensors [44, Ch.9], only an Identification error may be used. Again, the BI index should not be sensible to scale and permutation.

- When sources are non stationary [120], or have different spectra (even overlapping) [11], there exist separation techniques based on second order statistics. The latter technique, sometimes referred to as *SOBI*, is inspired from previous works [62] [45] [44, Ch.7].

- Taking again into account the time dimension, one can improve on performances. In speech for instance, not only sources are non stationary, but there are also often some silent periods. During these periods, the number of sources may be smaller than the number of sources, even if the total number of audio sources exceeds it. This allows to decompose *parsimonious* representations in *over-complete bases* [103] [100] [44, Ch.10]. This is actually linked to the concept of frames.

- If sources have a discrete distribution, specific techniques may be used, as we have already seen in these notes. But the subject is only shyly tackled in the literature (see [44, Ch.3] and references therein).

- *Convolutive mixtures* are widely studied for a dozen of years. This subject has been completed left aside for reasons of space, but references have been given in section 2.4 [44, Ch.8].

- *Semi-Blind* approaches assume that part of source samples are known by the receiver. If this information can be taken into account, the performances are expected to be better. This difficult subject is little addressed in the literature, despite its practical great interest [7] [17] [51] [153]. For instance, several mobile transmissions use this assumption (e.g. GSM); they are referred to as *semi-blind* techniques [44, Ch.15].

- To close this overview, one can insist on the lack of knowledge of tensors, and in particular on the unexpected topological properties of tensor spaces that have been above mentioned [43]. This is the subject of stuttering researches.

9 Exercises

Exercise 1 *Cumulant of independent variables.* Let $X = a\,Y + b\,Z$, where Y and Z are independent random variables with finite moments up to order r, and (a, b) deterministic. By using the definition of the second characteristic function, prove that rth order cumulants are related by: $\mathcal{C}_{x\,(r)} = a^r\,\mathcal{C}_{y\,(r)} + b^r\,\mathcal{C}_{z\,(r)}$

Exercise 2 *Real Gaussian Characteristic function.* Show that for a real Gaussian random variable x, the first characteristic function takes the form bellow, where $\mu_1' = \mathrm{E}\{x\}$ and $\mu_2 = \mathrm{E}\{(x - \mu_1')^2\}$:

$$\Phi(t) = \exp[\jmath\mu_1'\,t - \frac{1}{2}\,\mu_2\,t^2] \tag{75}$$

Exercise 3 *Cumulants may not exist.* Let the Cauchy distribution

$$p(u) = \frac{1}{\pi\,(1+u^2)}$$

Show that moments and cumulants of order larger than 1 are infinite. Explain why the mean is also undefined.

Exercise 4 *First cumulants.* Express the Cumulants of order 2 and 3 of a real scalar random variable as a function of moments. Assuming a zero mean and a symmetric distribution, express the 4th order cumulant as a function of moments.

Exercise 5 *Uniform distribution.* Show that, for a random variable uniformly distributed in the interval $[-u, +u]$, we have the following expressions for moments and cumulants: $\mu_{2k} = \frac{u^{2k}}{2k+1}$, $\mathcal{C}_4 = -\,2\,\frac{u^4}{15}$, $\mathcal{K}_4 = -\frac{6}{5}$.

Exercise 6 *Binary distribution.* Let x be a random variable taking two values x_1 and x_2 with probabilities P_1 and P_2. (i) Show that we must have $P_1 = \frac{1}{1+a^2}$, $P_2 = \frac{a^2}{1+a^2}$, and $x_1 = -a$, $x_2 = 1/a$, for some free parameter a, if x is zero-mean and unit variance. (ii) Show that the skewness is $\mathcal{C}_3 = 1/a - a$ and the kurtosis is $\mathcal{C}_4 = P_2\,[a^2 + 1/a^4] - 3$. (iii) What is the value of a and $\mathcal{C}_4$ in the symmetric case? What can be concluded?

Exercise 7 *Extremal values of the kurtosis.* Consider the set of probability densities defined on the real line with zero mean and unit variance. (i) What are the distributions reaching extreme values of the Nth order moment? (ii) What do we get for the kurtosis?

Exercise 8 *Circularity of PSK random variables.* PSK$-r$ random variables take their values in the set of rth roots of unity, with equal probabilities. Show that they are circular up to order $r - 1$.

Exercise 9 *The noncircular covariance.* The covariance matrix of a complex random variable is

$$\mathrm{E}\{z_i\,{z_j}^*\} - \mathrm{E}\{z_i\}\,\mathrm{E}\{z_j\}^* = \mathcal{C}_{z\,i}^{\;\;j}$$

On the other hand, if the random variable $\boldsymbol{z}$ is *circular at order 2*, then the moment $\mu_{z\,ij} = \mathrm{E}\{\boldsymbol{z}\,\boldsymbol{z}^{\mathsf{T}}\}_{ij}$ is null for all (i, j). Analyze the circularity at order 2 of the following random variables: (i) BPSK, (ii) PSK8, (iii) PAM4, (iv) QAM16. (v) what about the circularity of the 2-dimensional variable x, where x_1 is PSK8 and x_2 is PAM4, x_i being independent?

Exercise 10 *Strict circularity.* Show that a random variable z that is circular in the strict-sense is circular at any order.

Exercise 11 *Two sources without noise.* Assume a static model $\boldsymbol{x} = \boldsymbol{H}\,\boldsymbol{s}$ of dimension 2, where $\boldsymbol{s}$ has two statistically independent components. Also assume that a spatial pre-whitening has been performed, so that the standardized observation exactly follows the model below:

$$\tilde{\boldsymbol{x}} = \begin{pmatrix} \cos\alpha & -\sin\alpha\, e^{\jmath\varphi} \\ \sin\alpha\, e^{-\jmath\varphi} & \cos\alpha \end{pmatrix} \boldsymbol{s} \tag{76}$$

Denote $\gamma_{ij}^{k\ell} = \mathrm{Cum}\{\tilde{x}_i, \tilde{x}_j, \tilde{x}_k{}^*, \tilde{x}_\ell{}^*\}$ the fourth order circular cumulants of $\tilde{\boldsymbol{x}}$ and $\kappa_i = \mathrm{Cum}\{s_i, s_i, s_i{}^*, s_i{}^*\}$ those of s_i.
(i) Using the multi-linearity property of cumulants, establish the input-output relations between cross cumulants γ_{12}^{12}, γ_{11}^{12}, and γ_{12}^{22} on one hand, and κ_i, φ and α on the other hand.
(ii) From these expressions, deduce a closed-form solution for φ and α, and hence the mixing matrix.

Exercise 12 *Tendency to Gaussianity.* Let s_1 and s_2 be two statistically independent random variables, with positive kurtoses κ_1 and κ_2, respectively. Show that the kurtosis $\mathcal{K}$ of the random variable $x = a_1\, s_1 + a_2\, s_2$ is smaller than (or equal to) κ_i. *Hint:* without restricting the generality and for the sake of simplicity, assume that s_i have a unit variance, and show that $\mathcal{K}$ reaches its minimum for $\kappa_1\kappa_2/(\kappa_1+\kappa_2)$ and its maxima for κ_1 and κ_2.

NB: This result will be extended in two respects in this course: in the development of the cumulant-based Central Limit theorem, and in the construction of cumulant-based contrast criteria.

Exercise 13 *Inequality between skewness and kurtosis.* (i) Show the following inequality between *skewness* and *kurtosis* of real random variables:

$$\mathcal{K}_{(3)}{}^2 \leq \mathcal{K}_{(4)} + 2$$

One can restrict the proof to zero mean variables to make it simpler. (ii) Show that $-2 \leq \mathcal{K}_{(4)}$, and give an example for which the bound is reached. (iii) How could this result change in the complex case? *Hint:* consider the variance of a random variable of the form $aX + bX^2$, or the determinant of the Hankel matrix whose first row is $[1,\, \mu'_{x\,(1)},\, \mu'_{x\,(2)}]$ and last column $[\mu'_{x\,(2)},\, \mu'_{x\,(3)},\, \mu'_{x\,(4)}]^{\mathsf{T}}$.

Exercise 14 *Statistics of the Gauss distribution.* Let x be a real Gaussian random variable of probability distribution $p_x(u) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{u^2}{2\,\sigma^2})$. (i) Explain why odd order moments are null. (ii) Show that even order moments are given by

$$\mu_{x\,(2r)} = \sigma^{2r}\, \frac{(2r)!}{r!\, 2^r}. \tag{77}$$

(iii) Show that cumulants of order larger than 2 are all null. *Hint:* use the second characteristic function.

Exercise 15 *MI of a Gaussian variate.* Take a real Gaussian random variable of size K, of density

$$g_{\boldsymbol{x}}(\boldsymbol{u}) = [2\pi]^{-K/2}\, [\det \boldsymbol{V}]^{-1/2}\, e^{-\frac{1}{2}\boldsymbol{u}^{\mathsf{T}}\boldsymbol{V}^{-1}\boldsymbol{u}}, \tag{78}$$

where $\boldsymbol{V}$ is an invertible covariance matrix. Compute its Mutual Information. What happens if $\boldsymbol{V}$ is diagonal? Can this property be preserved under some congruent transforms?

Exercise 16 *Entropy of a Gaussian variate.* (i) Give the expression of the differential entropy of the Gaussian distribution (78), when the logarithm is computed in basis a. (ii) What happens if the variable is complex Gaussian circular?

Exercise 17 *Entropy of discrete variables.* Let a random variable X be equally distributed on N distinct symbols. What is its discrete entropy? If computed in basis 2, what would be for instance the entropy of a QAM64 random variable?

Exercise 18 *Extreme values of Entropy.* (i) What is the general form of distributions reaching extreme values of the differential entropy? What do we get for distributions with support: (ii) the whole real line $\mathbb{R}$, (iii) the positive real line $\mathbb{R}^+$, (iv) a finite interval $[a, b]$?

Exercise 19 *Shannon entropy.* Show that the Differential Entropy is not the limit of Shannon entropy, defined for discrete random variables.

Exercise 20 *Kullback divergence.* What is the Kullback divergence between a zero-mean uniform random variable and a zero-mean Gaussian variable of same variance?

Exercise 21 *Static* 2×2 *model.* In the noiseless real static 2×2 blind identification problem, one measures the following three cross cumulants: $\gamma_{1112} = \sqrt{3}/4$, $\gamma_{1222} = -\sqrt{3}/4$ et $\gamma_{1122} = -3/4$. Then what are the acceptable estimates of the mixing matrix among the following:

$$\begin{bmatrix} 1 & -\sqrt{3} \\ \sqrt{3} & 1 \end{bmatrix}, \begin{bmatrix} \sqrt{3} & 1 \\ -1 & \sqrt{3} \end{bmatrix}, \begin{bmatrix} \sqrt{3} & 1 \\ 1 & \sqrt{3} \end{bmatrix}, \begin{bmatrix} \sqrt{3}/2 & -\sqrt{3}/2 \\ 1/2 & 1/2 \end{bmatrix}$$

Can one estimate the source kurtoses?

Exercise 22 *Simplest counter-example of pairwise independence.* We are given a bag containing 4 Bowls: 1 Red, 1 Yellow, 1 Green, 1 with the 3 colors. Denote $\{RB, YB, GB, RYGB\}$ this set of bowls. The bowls have equal drawing probabilities $P(RB) = P(YB) = P(GB) = P(RYGB) = 1/4$. Define the event "$R$" of drawing a bowl containing the Red color.
(i) What is the probability of events R, Y, or G?
(ii) What is the probability of event $R \cap Y$? Are the events R, Y, G, pairwise independent?
(iii) What is the probability of $R \cap Y \cap G$? Are the events R, Y, G, mutually independent?

Exercise 23 *Pairwise independence of virtual sources.* Consider 3 mutually independent BPSK sources, $x_i \in \{-1, 1\}$, $1 \leq i \leq 3$. Define $x_4 = x_1 x_2 x_3$.
(i) Prove that x_4 is also BPSK, (but *dependent* on other x_i's).
(ii) Show that x_k are *pairwise independent*
(iii) Show that they are not *mutually independent. Hint:* one can show that that $\mathrm{Cum}\{x_1, x_2, x_3, x_4\} \neq 0$.

Exercise 24 *A class of general contrasts.* Let f be a convex strictly increasing function, with $f(0) = 0$, and let p and q be two positive (possibly null) integers so that $p + q > 2$. Denote $\zeta_{i\,(p)}^{(q)}$ the marginal cumulants at the output of a unitary mixture. Then show that the functional

$$\Upsilon_f(\boldsymbol{z}) \stackrel{\text{def}}{=} \sum_i f(|\zeta_{i\,(p)}^{(q)}|)$$

is a contrast, if used after prewhitening. *Hint:* for the domination property, use first the fact that f is increasing and that $\boldsymbol{G}$ is unitary, and then that f is convex.
Note: a real function f is convex iff for any set of positive numbers α_i such that $\sum_i \alpha_i = 1$, $f(\sum \alpha_i x_i) \leq \sum \alpha_i f(x_i)$.

Exercise 25 *Comparison of discrimination powers.* Let $\boldsymbol{x}$ be a random variable of size P and unit covariance matrix. It is desired to find a unitary matrix $\boldsymbol{Q}$ such that the random variable $\boldsymbol{z} = \boldsymbol{Q}\,\boldsymbol{x}$ maximizes a statistical contrast. Define the following contrast criteria:

$$\Upsilon_{CoM2}(\boldsymbol{Q}) = \sum_{i=1}^{P} |T_{iiii}|^2, \quad \Upsilon_{STD}(\boldsymbol{Q}) = \sum_{i=1}^{P}\sum_{j=1}^{P} |T_{iiij}|^2, \quad \Upsilon_{JAD}(\boldsymbol{Q}) = \sum_{i=1}^{P}\sum_{j=1}^{P}\sum_{k=1}^{P} |T_{iijk}|^2$$

where $T_{ijk\ell} = \mathrm{Cum}\{z_i,\, z_j,\, z_k,\, z_\ell\}$ denotes the fourth order cumulant of the output $\boldsymbol{z}$.
(i) Show that $\Upsilon_{CoM2}(\boldsymbol{Q}) \leq \Upsilon_{STD}(\boldsymbol{Q}) \leq \Upsilon_{JAD}(\boldsymbol{Q})$.
(ii) Show that, for any unitary matrix $\boldsymbol{Q}$, $\Upsilon_{JAD}(\boldsymbol{Q}) \leq \Upsilon_{CoM2}(\boldsymbol{I})$ with equality iff $\boldsymbol{Q} = \boldsymbol{\Lambda P}$ for some permutation $\boldsymbol{P}$ and diagonal matrix $\boldsymbol{\Lambda}$. What do you conclude?
(ii) Explain why these inequalities show that Υ_{CoM2} performs better than Υ_{STD}, itself performing better than Υ_{JAD}.

Exercise 26 *Sources with same kurtosis sign.* Assume all source processes have the same kurtosis sign, ε. Then show that the following is a contrast after prewhitening:

$$\Upsilon_{1,4} = \varepsilon \sum_{p=1}^{P} \zeta_{ii}^{ii}$$

Exercise 27 *Stationary points of contrast CoM2.* Let $\boldsymbol{T}$ be a real symmetric matrix, and let $\boldsymbol{G} = \boldsymbol{Q}\,\boldsymbol{T}\,\boldsymbol{Q}^{\mathsf{T}}$, where $\boldsymbol{Q}$ is a real orthogonal matrix. Define the CoM2 contrast function $\Upsilon_2(\boldsymbol{Q}) = \sum_i G_{ii}^2$.
(i) Show that the stationary points of $\Upsilon_2(\boldsymbol{Q})$ satisfy, for any pair of indices (q, r) with $q \neq r$: $G_{qq}\,G_{qr} = G_{rr}\,G_{qr}$.
(ii) Show that we have a maximum iff $G_{qr}^2 < (G_{qq} - G_{rr})^2$. What are are thus the possible extrema of $\Upsilon_2(\boldsymbol{Q})$?
(iii) Condiser now third order symmetric tensors $\boldsymbol{S}$ and $\boldsymbol{G}$, with $G_{ijk} = \sum_{a,b,c} Q_{ia}\,Q_{jb}\,Q_{kc}\,T_{abc}$. What are the stationary points of $\Upsilon_2(\boldsymbol{Q})$? Can we draw the same conclusions?
Hint: write the differential $d\boldsymbol{Q}$ as $d\boldsymbol{S}\,\boldsymbol{Q}$, where $d\boldsymbol{S}$ is skew symmetric, and use a canonical basis for the space of skew symmetric matrices.

Exercise 28 *Blind identification of a MA model.* Consider the SISO MA model $x[n] = \sum_{\ell=1}^{L} h[\ell]\,s[n-\ell] + v[n]$, where $s[n]$ is a i.i.d. non Gaussian sequence and $v[n]$ is Gaussian.
(i) By using the multi-linearity property of cumulants, show that the cumulant $C(i,j) \stackrel{\text{def}}{=} \mathrm{Cum}\{x[t],\, x[t+i],\, x[t+j],\, x[t+L]\}$ must be equal to $h[i]\,h[j]\,h[0]\,h[L]\,\kappa_s$, where κ_s denotes the 4th order cumulant of $s[n]$.
(ii) deduce a simple way of identifying the taps $h[i]$ from cumulant estimates $C(i,j)$ when the channel length L is known.

Exercise 29 *Complex rank:.* Define the $2 \times 2 \times 2$ tensor by its two matrix slices:

$$\boldsymbol{T}(:,:,1) = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \boldsymbol{T}(:,:,2) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

Check out that $\boldsymbol{T}$ can be decomposed in $\mathbb{R}$ as:

$$\boldsymbol{T} = \frac{1}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix}^{\circ 3} + \frac{1}{2}\begin{pmatrix} 1 \\ -1 \end{pmatrix}^{\circ 3} - 2\begin{pmatrix} 1 \\ 0 \end{pmatrix}^{\circ 3}$$

whereas it admits a CanD of rank 2 in $\mathbb{C}$:

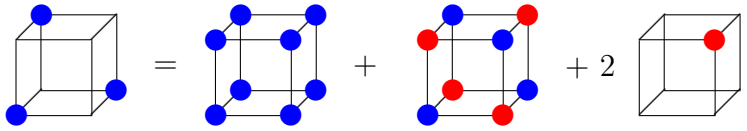
$$\boldsymbol{T} = \frac{\jmath}{2}\begin{pmatrix} -\jmath \\ 1 \end{pmatrix}^{\circ 3} - \frac{\jmath}{2}\begin{pmatrix} \jmath \\ 1 \end{pmatrix}^{\circ 3}$$

What do you conclude concerning the rank of $\boldsymbol{T}$?

Exercise 30 *Maximal rank of a tensor.* Define the $2 \times 2 \times 2$ symmetric tensor, null everywhere except for the entries $T_{112} = T_{121} = T_{211} = 1$. This tensor admits the decomposition:

$$\boldsymbol{T} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}^{\circ 3} + \begin{pmatrix} -1 \\ 1 \end{pmatrix}^{\circ 3} - 2\begin{pmatrix} 0 \\ 1 \end{pmatrix}^{\circ 3}$$

which is an explicit irreducible CAND. This decomposition is depicted in figure below.



blue bullets $= 1$, red bullets $= -1$.

Similarly, define the tensor of order 3 and dimension 3 by $T_{ijk} = 0$ everywhere except for the entries for which $i + j + k = 7$, for which the value is 1. It can be shown that this tensor has rank 5. Note that the typical rank is 4 in that case.

10 Bibliographical references

[1] K. Abed-Meraim, P. Loubaton, E. Moulines, "A subspace algorithm for certain blind identification problems", *IEEE Trans. Inf. Theory*, pp. 499–511, Mar. 1997.

[2] K. Abed-Meraim, E. Moulines, P. Loubaton, "Prediction error method for second-order blind identification", *IEEE Trans. Sig. Proc.*, vol. 45, no. 3, pp. 694–705, Mar. 1997.

[3] K. Abed-Meraim, Others, "On subspace methods for blind identification of SIMO FIR systems", *IEEE Trans. Sig. Proc.*, vol. 45, no. 1, pp. 42–55, Jan. 1997, Special issue on communications.

[4] L. Albera, A. Ferreol, P. Comon, P. Chevalier, "Blind identification of overcomplete mixtures of sources (BIOME)", *Lin. Algebra Appl.*, vol. 391, pp. 1–30, Nov. 2004.

[5] S. Amari, "Natural gradient works efficiently in learning", *Neural Computation*, vol. 10, pp. 251–276, 1998.

[6] S. Amari, H. Nagaoka, *Methods of Information Geometry*, Am. Math. Soc., 2000.

[7] J. Ayadi, E. Decarvalho, D. T. M. Slock, "Blind and semi-blind maximum likelihood methods for FIR multichannel identification", in *Proc. ICASSP*, Seattle, May 12-15 1998.

[8] M. Babaie-Zadeh, C. Jutten, K. Nayebi, "Blind separation of post-nonlinear mixtures", in *Int. Conf. Indep. Comp. Ana. (ICA'01)*, San Diego, Dec. 2001, pp. 138–143.

[9] Y. Bar-Ness, J. W. Carlin, M. L. Steinberger, "Bootstrapping adaptive interference cancelers: Some practical limitations", in *Proc. The Globecom. Conference*, Miami, Nov. 1982, pp. 1251–1255, paper No F3. 7.

[10] A. J. Bell, T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution", *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, Nov. 1995.

[11] A. Belouchrani, K. Abed-Meraim, J. F. Cardoso, E. MOULINES, "A blind source separation technique using second order statistics", *IEEE Trans. Sig. Proc.*, vol. 45, no. 2, pp. 434–444, Feb. 1997.

[12] A. Benveniste, M. Goursat, G. Ruget, "Robust identification of a non-minimum phase system", *IEEE Trans. Auto. Contr.*, vol. 25, no. 3, pp. 385–399, June 1980.

[13] J. M. F. Ten Berge, N. Sidiropoulos, "On uniqueness in Candecomp/Parafac", *Psychometrika*, vol. 67, no. 3, pp. 399–409, Sept. 2002.

[14] J. Brachat, P. Comon, B. Mourrain, E. Tsigaridas, "Symmetric tensor decomposition", *Linear Algebra Appl.*, 2010, accepted. hal:inria-00355713, arXiv:0901.3706.

[15] R. Bro, "Parafac, tutorial and applications", *Chemom. Intel. Lab. Syst.*, vol. 38, pp. 149–171, 1997.

[16] H. Broman, U. Lindgren, H. Sahlin, P Stoica, "Source separation: a TITO system identification approach", *Signal Processing, Elsevier*, vol. 73, no. 2, Feb. 1999, special issue on blind separation and deconvolution.

[17] V. Buchoux, O. Cappe, Others, "On the performance of semi-blind subspace-based channel estimation", *Trans. on Sig. Proc.*, vol. 48, no. 6, pp. 1750–1759, June 2000.

[18] A. A. Cadzow, "Blind deconvolution via cumulant extrema", *IEEE Sig. Proc. Mag.*, vol. 13, pp. 24–42, May 1996.

[19] J. F. Cardoso, "Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem", in *Proc. ICASSP* Albuquerque, 1990, pp. 2655–2658.

[20] J. F. Cardoso, "Super-symmetric decomposition of the fourth-order cumulant tensor. Blind identification of more sources than sensors", in *Proc. ICASSP*, Toronto, 1991, pp. 3109–3112.

[21] J. F. Cardoso, "Iterative techniques for blind source separation using only fourth order cumulants", in *Proc. EUSIPCO*, Brussels, Belgium, 1992.

[22] J. F. Cardoso, "High-order contrasts for independent component analysis", *Neural Computation*, vol. 11, no. 1, pp. 157–192, Jan. 1999.

[23] J. F. Cardoso, A. Souloumiac, "Blind beamforming for non-Gaussian signals", *IEE Proceedings - Part F*, vol. 140, no. 6, pp. 362–370, Dec. 1993, Special issue on Applications of High-Order Statistics.

[24] J. D. Carroll, J. J. Chang, "Analysis of individual differences in multidimensional scaling via n-way generalization of Eckart-Young decomposition", *Psychometrika*, vol. 35, no. 3, pp. 283–319, Sept. 1970.

[25] P. Chevalier, "Méthodes aveugles de filtrage d'antennes", *Revue d'Electronique et d'Electricité, SEE*, , no. 3, pp. 48–58, Sept. 1995.

[26] P. Chevalier, "Optimal array processing for non stationary signals", in *ICASSP*, Atlanta, May 1996, vol. 5, pp. 2868–2871.

[27] P. Chevalier, L. Albera, A. Ferreol, P. Comon, "On the virtual array concept for higher order array processing", *IEEE Trans. Sig. Proc.*, vol. 53, no. 4, pp. 1254–1271, Apr. 2005.

[28] P. Chevalier, P. Comon, "Séparation aveugle de sources", in *Traité des Télécommunications*, pp. 1–22. Techniques de l'Ingénieur, Paris, 2002, TE 5 250.

[29] A. Chevreuil, P. Loubaton, "Modulation and second order cyclostationarity: Structured subspace method and new identifiability conditions", *IEEE Sig. Proc. Letters*, vol. 4, no. 7, pp. 204–206, July 1997.

[30] A. Cichocki, S-I. Amari, *Adaptive Blind Signal and Image Processing*, Wiley, New York, 2002.

[31] P. Comon, "Separation of stochastic processes", in *Proc. Workshop on Higher-Order Spectral Analysis*, Vail, Colorado, June 28-30 1989, IEEE-ONR-NSF, pp. 174–179.

[32] P. Comon, "Analyse en Composantes Indépendantes et identification aveugle", *Traitement du Signal*, vol. 7, no. 3, pp. 435–450, Dec. 1990, Numero special non lineaire et non gaussien.

[33] P. Comon, "Independent Component Analysis", in *Proc. Int. Sig. Proc. Workshop on Higher-Order Statistics*, Chamrousse, France, July 10-12 1991, pp. 111–120, Keynote address. Republished in *Higher-Order Statistics*, J.L.Lacoume ed., Elsevier, 1992, pp 29–38.

[34] P. Comon, "MA identification using fourth order cumulants", *Signal Processing, Elsevier*, vol. 26, no. 3, pp. 381–388, 1992.

[35] P. Comon, "Independent Component Analysis, a new concept ?", *Signal Processing, Elsevier*, vol. 36, no. 3, pp. 287–314, Apr. 1994, Special issue on Higher-Order Statistics.

[36] P. Comon, "Tensor diagonalization, a useful tool in signal processing", in *IFAC-SYSID, 10th IFAC Symposium on System Identification*, M. Blanke, T. Soderstrom, Eds., Copenhagen, Denmark, July 4-6 1994, vol. 1, pp. 77–82, invited session.

[37] P. Comon, "Contrasts for multichannel blind deconvolution", *IEEE Signal Processing Letters*, vol. 3, no. 7, pp. 209–211, July 1996.

[38] P. Comon, "From source separation to blind equalization, contrast-based approaches", in *Int. Conf. on Image and Signal Processing (ICISP'01)*, Agadir, Morocco, May 3-5, 2001, invited plenary.

[39] P. Comon, "Tensor decompositions, state of the art and applications", in *Mathematics in Signal Processing V*, J. G. McWhirter, I. K. Proudler, Eds., pp. 1–24. Clarendon Press, Oxford, UK, 2002, arXiv:0905.0454v1.

[40] P. Comon, "Blind identification and source separation in 2x3 under-determined mixtures", *IEEE Trans. Signal Processing*, vol. 52, no. 1, pp. 11–22, Jan. 2004.

[41] P. Comon, "Contrasts, independent component analysis, and blind deconvolution", *Int. Journal Adapt. Control Sig. Proc.*, vol. 18, no. 3, pp. 225–243, Apr. 2004, special issue on Signal Separation: http://www3.interscience.wiley.com/cgi-bin/jhome/4508. Preprint: I3S Research Report RR-2003-06.

[42] P. Comon, P. Chevalier, "Source separation: Models, concepts, algorithms and performance", in *Unsupervised Adaptive Filtering, Vol. I, Blind Source Separation*, S. Haykin, Ed., Series on Adaptive and learning systems for communications signal processing and control, pp. 191–236. Wiley, 2000.

[43] P. Comon, G. Golub, L-H. Lim, B. Mourrain, "Symmetric tensors and symmetric tensor rank", *SIAM Journal on Matrix Analysis Appl.*, vol. 30, no. 3, pp. 1254–1279, Sept. 2008.

[44] P. Comon, C. Jutten, *Handbook of Blind Source Separation, Independent Component Analysis and Applications*, Academic Press, 2010.

[45] P. Comon, J. L. Lacoume, "Statistiques d'ordres supérieurs pour le traitement du signal", Ecole Predoctorale de Physique, Les Houches, 30 aout – 10 septembre 1993, P. Flandrin et J. L. Lacoume ed.

[46] P. Comon, X. Luciani, A. L. F. De Almeida, "Tensor decompositions, alternating least squares and other tales", *Jour. Chemometrics*, vol. 23, pp. 393–405, Aug. 2009.

[47] P. Comon, E. Moreau, "Improved contrast dedicated to blind separation in communications", in *ICASSP*, Munich, April 20-24 1997, pp. 3453–3456.

[48] P. Comon, B. Mourrain, "Decomposition of quantics in sums of powers of linear forms", *Signal Processing, Elsevier*, vol. 53, no. 2, pp. 93–107, Sept. 1996, special issue on High-Order Statistics.

[49] P. Comon, M. Rajih, "Blind identification of under-determined mixtures based on the characteristic function", in *ICASSP'05*, Philadelphia, March 18–23 2005, vol. IV, pp. 1005–1008.

[50] G. Darmois, "Analyse générale des liaisons stochastiques", *Rev. Inst. Internat. Stoch.*, vol. 21, pp. 2–8, 1953.

[51] Decarvalho, D. T. M. Slock, "Cramer-Rao bounds for semi-blind, blind and training sequence based channel estimation", in *Proc. SPAWC 97 Conf.*, Paris, France, Apr. 1997, pp. 129–132.

[52] N. Delfosse, P. Loubaton, "Adaptive blind separation of independent sources: a deflation approach", *Signal Processing*, vol. 45, pp. 59–83, 1995.

[53] F. Desbouvries, *Problemes Structures et Algorithmes du Second Ordre en Traitement du Signal*, HdR Marne la vallée, 22 Jun 2001.

[54] . Ding, "An outer-product decomposition algorithm for multichannel blind identification", in *8th IEEE SP Workshop SSAP*, Corfu, Greece, June 1996, pp. 132–135.

[55] Z. Ding, "Matrix outer-product decomposition method for blind multiple channel identification", *IEEE Trans. Sig. Proc.*, vol. 45, no. 12, pp. 3053–3061, Dec. 1997.

[56] Z. Ding, Y. Li, *Blind Equalization and Identification*, Dekker, New York, 2001.

[57] D. Donoho, "On minimum entropy deconvolution", in *Applied time-series analysis II*, pp. 565–609. Academic Press, 1981.

[58] D. Dugué, "Analycité et convexité des fonctions caractéristiques", *Annales de l'Institut Henri poincarré*, vol. XII, pp. 45–56, 1951.

[59] J. Eriksson, V. Koivunen, "Identifiability, separability and uniquness of linear ICA models", *IEEE Sig. Proc. Letters*, pp. 601–604, July 2004.

[60] W. Feller, *An Introduction to Probability Theory and its Applications*, vol. II, Wiley, 1966, 1971, 2nd ed.

[61] A. Ferreol, P. Chevalier, "On the behavior of current second and higher order blind source separation methods for cyclostationary sources", *IEEE Trans. Sig. Proc.*, vol. 48, pp. 1712–1725, June 2000, erratum in vol. 50, pp. 990, Apr. 2002.

[62] L. Fety, *Methodes de Traitement d'Antenne Adaptees aux Radiocommunications*, Doctorat, ENST, 1988.

[63] . D. Forney, "Minimal bases of rational vector spaces, with applications to multivariable linear systems", *SIAM J. Contr.*, vol. 13, pp. 493–520, May 1975.

[64] F. Gamboa, E. Gassiat, "Blind deconvolution of discrete linear systems", *Annals of Stat.*, Nov. 1996.

[65] E. Gassiat, F. Gamboa, "Source separation when the input sources are discrete or have constant modulus", *IEEE Trans. Sig. Proc.*, vol. 45, no. 12, pp. 3062–3072, Dec. 1997.

[66] D. Gesbert, P. Duhamel, S. Mayrargue, "On-line blind multichannel equalization based on mutually referenced filters", *IEEE Trans. Sig. Proc.*, vol. 45, no. 9, pp. 2307–2317, Sept. 1997.

[67] D. Godard, "Self recovering equalization and carrier tracking in two dimensional data communication systems", *IEEE Trans. Com.*, vol. 28, no. 11, pp. 1867–1875, Nov. 1980.

[68] A. Gorokhov, P. Loubaton, "Subspace based techniques for blind separation of convolutive mixtures with temporally correlated sources", *IEEE Trans. Cir. Syst.*, vol. 44, pp. 813–820, Sept. 1997.

[69] A. Gorokhov, P. Loubaton, "Blind identification of MIMO-FIR systems: a generalized linear prediction approach", *Signal Processing, Elsevier*, vol. 73, pp. 105–124, Feb. 1999.

[70] O. Grellier, *Deconvolution et Separation de Sources Discretes*, Doctorat UNSA, 26 Janvier 2000.

[71] O. Grellier, P. Comon, "Blind separation and equalization of a channel with MSK inputs", in *SPIE Conference*, San Diego, July 19-24 1998, pp. 26–34, invited session.

[72] O. Grellier, P. Comon, "Blind separation of discrete sources", *IEEE Signal Processing Letters*, vol. 5, no. 8, pp. 212–214, Aug. 1998.

[73] O. Grellier, P. Comon, B. Mourrain, P. Trebuchet, "Analytical blind channel identification", *IEEE Trans. Signal Processing*, vol. 50, no. 9, pp. 2196–2207, Sept. 2002.

[74] M. I. Gürelli, C. L. Nikias, "EVAM: An eigenvector -based algorithm for multichannel blind deconvolution of input colored signals", *IEEE Trans. Sig. Proc.*, vol. 43, no. 1, pp. 134–149, Jan. 1995.

[75] R. A. Harshman, "Foundations of the Parafac procedure: Models and conditions for an explanatory multimodal factor analysis", *UCLA Working Papers in Phonetics*, vol. 16, pp. 1–84, 1970, http://publish. uwo. ca/ harshman.

[76] S. Haykin, *Unsupervised Adaptive Filtering*, vol. 1, Wiley, 2000, series in Adaptive and Learning Systems for Communications, Signal Processing, and Control.

[77] F. L. Hitchcock, "The expression of a tensor or a polyadic as a sum of products", *J. Math. and Phys.*, vol. 6, no. 1, pp. 165–189, 1927.

[78] Y. Hua, "Fast maximum likelihood for blind identification of multiple fir channels", *IEEE Trans. Sig. Proc.*, vol. 44, no. 3, pp. 661–672, Mar. 1996.

[79] P. J. Huber, "Projection pursuit", *The Annals of Statistics*, vol. 13, no. 2, pp. 435–475, 1985, Invited paper with discussion.

[80] A. Hyvärinen, "A family of fixed-point algorithms for Independent Component Analysis", in *ICASSP'97*, Munich, Germany, April 20-24 1997, pp. 3917–3920.

[81] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis", *IEEE Trans. Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[82] A. Hyvärinen, J. Karhunen, E. Oja, *Independent Component Analysis*, Wiley, 2001.

[83] T. Jiang, N. Sidiropoulos, "Kruskal's permutation lemma and the identification of CANDECOMP/PARAFAC and bilinear models", *IEEE Trans. Sig. Proc.*, vol. 52, no. 9, pp. 2625–2636, Sept. 2004.

[84] C. Jutten, J. Hérault, "Independent component analysis versus PCA", in *Proc. EUSIPCO,* Grenoble, France, 1988, pp. 643–646.

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

[86] T. Kailath, *Linear Systems*, Prentice-Hall, 1980.

[87] M. Kendall, A. Stuart, *The Advanced Theory of Statistics, Distribution Theory*, vol. 1, C. Griffin, 1977.

[88] P. M. Kroonenberg, J. De Leeuw, "Principal component analysis of three-mode data", *Psychometrika*, vol. 45, pp. 69–97, 1980.

[89] J. B. Kruskal, "Three-way arrays: Rank and uniqueness of trilinear decompositions", *Linear Algebra and Applications*, vol. 18, pp. 95–138, 1977.

[90] J. L. Lacoume, P. O. Amblard, P. Comon, *Statistiques d'ordre supérieur pour le traitement du signal*, Collection Sciences de l'Ingénieur. Masson, 1997.

[91] L. De Lathauwer, *Signal Processing based on Multilinear Algebra*, Doctorate, Katholieke Universiteit Leuven, Sept. 1997.

[92] L. De Lathauwer, B. De Moor, "From matrix to tensor : Multilinear algebra and signal processing", in *Mathematics in Signal Processing IV*, Oxford, UK, 1998, pp. 1–15, Oxford University Press, selected papers presented at 4th IMA Int. Conf. on Math. Sig. Proc.

[93] L. De Lathauwer, B. De Moor, J. Vandewalle, "A singular value decomposition for higher-order tensors", in *Second ATHOS workshop*, Sophia-Antipolis, France, Sept 20-21 1993.

[94] L. De Lathauwer, B. De Moor, J. Vandewalle, "Fetal electrocardiogram extraction by blind source subspace separation", *IEEE Trans. Biomedical Engineering*, vol. 47, no. 5, pp. 567–572, May 2000, Special Topic Section on Advances in Statistical Signal Processing for Biomedicine.

[95] L. De Lathauwer, B. De Moor, J. Vandewalle, "A multilinear singular value decomposition", *SIAM Jour. Matrix Ana. Appl.*, vol. 21, no. 4, pp. 1253–1278, Apr. 2000.

[96] L. De Lathauwer, B. De Moor, J. Vandewalle, "Independent Component Analysis and (simultaneous) third-order tensor diagonalization", *IEEE Trans. Sig. Proc.*, pp. 2262–2271, Oct. 2001.

[97] J. Lebrun, P. Comon, "An algebraic approach to blind identification of communication channels", in *IEEE ISSPA'03*, Paris, 1-4 July 2003.

[98] J. Lebrun, P. Comon, "A linear algebra approach to systems of polynomial equations with application to digital communications", in *Eusipco'04*, Wien, Austria, Sep 7–10 2004.

[99] T. W. Lee, *Independent Component Analysis*, Kluwer, 1998.

[100] T. W. Lee, M. S. Lewicki, Others, "Blind source separation of more sources than mixtures using overcomplete representations", *IEEE Sig. Proc. Letters*, vol. 6, no. 4, pp. 87–90, Apr. 1999.

[101] J. Leroux, P. Sole, "Least-square error reconstruction of a sampled signal Fourier transform from its n-th order polyspectrum", *Signal Processing*, vol. 35, pp. 75–81, 1994.

[102] J. Levin, "Three-mode factor analysis", *Psychological Bulletin*, vol. 64, pp. 442–452, 1965.

[103] M. Lewicki, T. J. Sejnowski, "Learning non-linear overcomplete representations for efficient coding", in *Advances in Neural Information Processing Systems*, 1998, pp. 815–821.

[104] T. H. Li, "Analysis of a non-parametric blind equalizer for discrete valued signals", *IEEE Trans. on Sig. Proc.*, vol. 47, no. 4, pp. 925–935, Apr 1999.

[105] T. H. Li, K. Mbarek, "A blind equalizer for nonstationary discrete-valued signals", *IEEE Trans. Sig. Proc.*, vol. 45, no. 1, pp. 247–254, Jan. 1997, Special issue on communications.

[106] P. Liavas, P. A. Regalia, J-P. Delmas, "On the robustness of the linear prediction method for blind channel identification", *IEEE Trans. Sig. Proc.*, vol. 48, pp. 1477–1481, May 2000.

[107] P. Loubaton, E. Moulines, "On blind multiuser forward link channel estimation by subspace method: Identifiability results", *IEEE Trans. Sig. Proc.*, vol. 48, no. 8, pp. 2366–2376, Aug. 2000.

[108] P. Loubaton, E. Moulines, P. A. Regalia, "Subspace methods for blind identification and deconvolution", in *Signal Processing Advances in Wireless and Mobile Communications*, Giannakis, Hua, Stoica, Tong, Eds., chapter 3. Prentice-Hall, 2001.

[109] G. Lukacs, *Characteristic functions*, Griffin, 1960.

[110] A. Mansour, C. Jutten, N. Ohnishi, "Kurtosis: Definition and properties", in *International Conference on Multisource-Multisensor: Data Fusion (FUSION'98)*, Las Vegas, USA, 6-9 July 1998, pp. 40–46.

[111] J. C. Marron, P. P. Sanchez, R. C. Sullivan, "Unwrapping algorithm for least-squares phase recovery from the modulo-2π bispectrum phase", *Jour. Opt. Soc. Amer.*, vol. 7, pp. 14–20, 1990.

[112] T. Matsuoka, T. J. Ulrych, "Phase estimation using the bispectrum", *Proc. IEEE*, vol. 72, 1984.

[113] P. Mccullagh, *Tensor Methods in Statistics*, Monographs on Statistics and Applied Probability. Chapman and Hall, 1987.

[114] E. Moreau, "A generalization of joint-diagonalization criteria for source separation", *IEEE Trans. Signal Processing*, vol. 49, no. 3, pp. 530–541, March 2001.

[115] E. Moulines, J. F. Cardoso, "Second-order versus fourth-order MUSIC algorithms. an asymptotical statistical performance analysis", in *Proc. Int. Sig. Proc. Workshop on Higher-Order Statistics*, Chamrousse, France, 1991, pp. 121–130.

[116] E. Moulines, P. Duhamel, J. F. Cardoso, S. MAYRAGUE, "Subspace methods for the blind identification of multichannel FIR filters", *IEEE Trans. Sig. Proc.*, vol. 43, no. 2, pp. 516–525, Feb. 1995.

[117] C. L. Nikias, A. P. Petropulu, *Higher-Order Spectra Analysis*, Signal Processing Series. Prentice-Hall, Englewood Cliffs, 1993.

[118] D. T. Pham, "Blind separation of instantaneous mixture of sources via an independent component analysis", *IEEE Trans. Sig. Proc.*, vol. 44, no. 11, pp. 2768–2779, Nov. 1996.

[119] D. T. Pham, "Contrast functions for blind separation and deconvolution of sources", in *Int. Conf. Indep. Comp. Ana. (ICA'01)*, San Diego, Dec. 2001.

[120] D. T. Pham, J-F. Cardoso, "Blind separation of instantaneous mixtures of nonstationary sources", *IEEE Trans. Sig. Proc.*, vol. 49, no. 9, pp. 1837–1848, Sept. 2001.

[121] D. T. Pham, P. Garat, "Blind separation of mixture of independent sources through a quasi-maximum likelihood approach", *IEEE Trans. Sig. Proc.*, vol. 45, no. 7, pp. 1712–1725, July 1997.

[122] J. G. Proakis, *Digital Communications*, McGraw-Hill, 1995, 3rd edition.

[123] G. Salmon, *Lessons introductory to the Modern Higher Algebra*, Chesla publ., New York, 1885.

[124] Y. Sato, "A method of self recovering equalization for multilevel amplitude-modulation systems", *IEEE Trans. Com.*, vol. 23, pp. 679–682, June 1975.

[125] O. Shalvi, E. Weinstein, "New criteria for blind deconvolution of nonminimum phase systems", *IEEE Trans. Inf. Theory*, vol. 36, no. 2, pp. 312–321, Mar. 1990.

[126] J. E. Shore, R. W. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy", *IEEE Trans. Information Theory*, vol. 26, no. 1, pp. 26–37, Jan. 1980.

[127] N. D. Sidiropoulos, R. Bro, "On the uniqueness of multilinear decomposition of N-way arrays", *Jour. Chemo.*, vol. 14, pp. 229–239, 2000.

[128] N. D. Sidiropoulos, R. Bro, G. B. Giannakis, "Parallel factor analysis in sensor array processing", *IEEE Trans. Sig. Proc.*, vol. 48, no. 8, pp. 2377–2388, Aug. 2000.

[129] C. Simon, P. Loubaton, C. Jutten, "Separation of a class of convolutive mixtures: A contrast function approach", in *ICASSP*, Phoenix, May 1999.

[130] C. Simon, P. Loubaton, C. Vignat, C. Jutten, G. D'urso, "Blind source separation of convolutive mixtures by maximization of fourth-order cumulants: the non i. i. d. case", in *ICASSP*, Seattle, May 12-15 1998.

[131] T. M. Slock, "Blind fractionally-spaced equalization, perfect-reconstruction filter banks and multichannel linear prediction", in *Proc. ICASSP 94 Conf.*, Adelaide, Australia, April 1994.

[132] A. Smilde, R. Bro, P. Geladi, *Multi-Way Analysis*, Wiley, 2004.

[133] A. Swami, G. Giannakis, S. Shamsunder, "Multichannel ARMA processes", *IEEE Trans. Sig. Proc.*, vol. 42, no. 4, pp. 898–913, Apr. 1994.

[134] A. L. Swindlehurst, S. Daas, J. Yang, "Analysis of a decision directed beamformer", *IEEE Trans. Sig. Proc.*, vol. 43, no. 12, pp. 2920–2927, Dec. 1995.

[135] A. Taleb, C. Jutten, "Non-linear source separation: the post-non-linear mixtures", in *ESANN'97*, Bruges, Belgium, 1997, pp. 279–284.

[136] A. Taleb, C. Jutten, "Source separation in post-nonlinear mixtures", *IEEE Trans. Sig. Proc.*, vol. 47, pp. 2807–2820, Oct. 1999.

[137] S. Talwar, M. Viberg, A. Paulraj, "Blind estimation of multiple co-channel digital signals arriving at an antenna array: Part I, algorithms", *IEEE Trans. Sig. Proc.*, pp. 1184–1197, May 1996.

[138] L. Tong, "Identification of multichannel MA parameters using higher-order statistics", *Signal Processing, Elsevier*, vol. 53, no. 2, pp. 195–209, Sept. 1996, special issue on High-Order Statistics.

[139] L. Tong, G. Xu, T. Kailath, "Blind identification and equalization based on second-order statistics: a time domain approach", *IEEE Trans. Inf. Theory*, vol. 40, no. 2, pp. 340–349, Mar. 1994.

[140] A. Touzni, *Performance et Robustesse en Egalisation Spatio-Temporelle*, Doctorat de Cergy-Pontoise, 1998.

[141] A. Touzni, I. Fijalkow, M. Larimore, J. R. Treichler, "A globally convergent approach for blind MIMO adaptive deconvolution", in *ICASSP*, Seattle, May 12-15 1998.

[142] J. R. Treichler, M. G. Larimore, "New processing techniques based on the constant modulus adaptive algorithm", *IEEE Trans. Acoust. Speech Sig. Proc.*, vol. 33, no. 2, pp. 420–431, Apr. 1985.

[143] L. R. Tucker, "Some mathematical notes for three-mode factor analysis", *Psychometrika*, vol. 31, pp. 279–311, 1966.

[144] J. K. Tugnait, "Approaches to FIR system identification with noisy data using higher-order statistics", *IEEE Trans. Acoust. Speech Sig. Proc.*, pp. 1307–1317, July 1990.

[145] J. K. Tugnait, "Blind spatio-temporal equalization and impulse response estimation for MIMO channels using a Godard cost function", *IEEE Trans. Sig. Proc.*, vol. 45, no. 1, pp. 268–271, Jan. 1997.

[146] J. K. Tugnait, L. Tong, Z. Ding, "Single-user channel estimation and equalization", *IEEE Signal Processing Magazine*, vol. 17, no. 3, pp. 17–28, May 2000.

[147] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice-Hall, London, 1993.

[148] A. J. Van Der Veen, A. Paulraj, "An analytical constant modulus algorithm", *IEEE Trans. Sig. Proc.*, vol. 44, no. 5, pp. 1136–1155, May 1996.

[149] A. J. Van Der Veen, S. Talwar, A. Paulraj, "Blind estimation of multiple digital signals transmitted over FIR channels", *IEEE Sig. Proc. Letters*, vol. 2, no. 5, pp. 99–102, May 1995.

[150] T. Wigren, "Avoiding ill-convergence of finite dimensional blind adaptation schemes excited by discrete symbol sequences", *Signal Processing, Elsevier*, vol. 62, no. 2, pp. 121–162, Oct. 1997.

[151] G. Xu, H. Liu, L. Tong, T. Kailath, "A least-squares approach to blind channel identification", *IEEE Trans. Sig. Proc.*, vol. 43, no. 12, pp. 813–817, Dec. 1995.

[152] D. Yellin, B. Porat, "Blind identification of FIR systems excited by discrete-alphabet inputs", *IEEE Trans. Sig. Proc.*, vol. 41, no. 3, pp. 1331–1339, 1993.

[153] V. Zarzoso, P. Comon, "Blind and semi-blind equalization based on the constant power criterion", *IEEE Trans. Sig. Proc.*, vol. 53, no. 11, pp. 4363–4375, Nov. 2005.

[154] V. Zivojnovic, "Higher-order statistics and Huber's robustness", in *IEEE-ATHOS Workshop on Higher-Order Statistics*, Begur, Spain, 12–14 June 1995, pp. 236–240.

Index