TUT#10

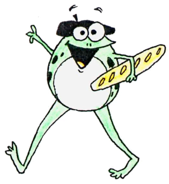
Independent Component Analysis & Multi-way Factor Analysis

Pierre Comon

Lab. I3S
CNRS & University of Nice
Sophia-Antipolis, France

Contents

Part I: General



- Goal of ICA, example
- Uniqueness
- Applications
- Decorrelation vs Independence
- Taxonomy
- Brief historical survey

Principal Component Analysis (PCA)

Goal

Given a K-dimensional r.v., $\boldsymbol{x}$, find $\boldsymbol{U}$ and $\boldsymbol{z}$ such that

- Observation

$$\boldsymbol{x} = \boldsymbol{U}\,\boldsymbol{z}$$

- $\boldsymbol{z}$ has uncorrelated components z_i

NB: Because of lack of uniqueness, $\boldsymbol{U}$ is often assumed to be unitary.

General

Independent Component Analysis (ICA)

Goal

Given a K-dimensional r.v., $\boldsymbol{x}$, find $\boldsymbol{H}$ and $\boldsymbol{s}$ such that

- Observation

$$\boldsymbol{x} = \boldsymbol{H}\,\boldsymbol{s} \qquad (1)$$

- $\boldsymbol{s}$ has mutually statistically independent components s_i

➠ *"Blind"* terminology: only outputs x_i are observed.

General

Uniqueness

Inherent indeterminations

if $\boldsymbol{s}$ has independent components s_i, so has $\boldsymbol{\Lambda P}\,\boldsymbol{s}$

where $\boldsymbol{\Lambda}$ is invertible diagonal and $\boldsymbol{P}$ permutation

Solutions

If $(\boldsymbol{A},\,\boldsymbol{s})$ solution, then $(\boldsymbol{A\Lambda P},\,\boldsymbol{P}^{\mathsf{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{s})$ also is.

- *"Essential uniqueness"*: unique up to a *trivial filter*, i.e. a scale-permutation
- Whole equivalence class of solutions ⇒ Look for one representative.

General

Decorrelation vs Independence

Example 1: Mixture of 2 identically distributed sources

Consider the mixture of two independent sources

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} s_1 \\ s_2 \end{pmatrix}$$

where $\mathrm{E}\{s_i^2\} = 1$ and $\mathrm{E}\{s_i\} = 0$. Then x_i are *uncorrelated*:

$$\mathrm{E}\{x_1\, x_2\} = \mathrm{E}\{s_1^2\} - \mathrm{E}\{s_2^2\} = 0$$

But x_i are *not independent* since, for instance:

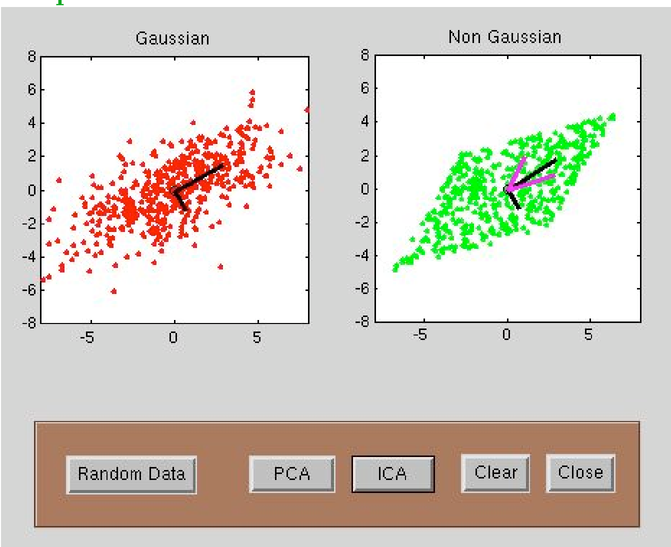
$$\mathrm{E}\{x_1^2\, x_2^2\} - \mathrm{E}\{x_1^2\}\mathrm{E}\{x_2^2\} = \mathrm{E}\{s_1^4\} + \mathrm{E}\{s_2^4\} - 6 \neq 0$$

▷ demoICA2x2

General

PCA vs ICA

Example 2: 2 sources and 2 sensors



General

Applications (1)

- Sensor Array Processing
  - Localization with reduced diversity
  - Localization with ill calibrated antennas
  - Detection and/or extraction with unknown antennas (eg. sonar buoys, biomedical, audio, nuclear plants...)
  - Blind extraction (eg. ComInt: interception, surveillance)

General

Applications (2)

- Factor Analysis
  - Chemometrics
  - Econometrics
  - Psychology
- Denoising
- Compression
- Arithmetic Complexity
- Machine Learning
- Exploratory Analysis

General

Taxonomy (1)

Static/Dynamic and **Noisy/Noiseless**:

$$\boldsymbol{x}[n] = \boldsymbol{H} \star \boldsymbol{s}[n] + \boldsymbol{v}[n] \quad (2)$$

Linearly **Invertible/Under-Determined:**

$$Number\ of\ sources:\ P \underset{\text{Invertible}}{\overset{\text{UnderDet}}{\gtrless}} K\ :\ Number\ of\ sensors$$

General

Taxonomy (2)

Transmit/Receive diversity:

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

General

Taxonomy (3)

Assumptions required on sources:

- H1. use of Time coherency of $\boldsymbol{s}(n)$: separation by exploiting spectral differences.
- H2. Sources s_i are mutually statistically independent
  - Static case: r.v. statistically independent (but may have identical p.s.d.) $\longrightarrow$ ICA
  - Dynamic case: Sources are i.i.d. (i.e. white) processes
- H3. Sources are Discrete (but may be stat. dependent)
- H4. Sources are non stationary (and have different time profiles)

General

Historical survey (Static MIMO only)

- **The ancestors:** Dugué'51, Darmois'53, Feller'66, Friedman'74, Donoho'80
- **The first shy steps in ICA:** Bar-Ness'82, Jutten'83, Fety'88
- **The first steps ins Multi-way:** Carroll-Chang'70, Harshman'70, Kruskal'77
- **First closed-form solutions:** Comon'89, Cardoso'92
- **First IT frameworks:** Comon'91, Cardoso'93, Comon'94, Bell'95, Delfosse-Loubaton'95
- **Specific improvements:** Hyvarinen'97, Pajunen'97, Amari'98, Grellier'98, Parra'2000
- **Recent advances:** Cao-Liu'96, Moreau-Pesquet'97, Taleb-Jutten'97, Comon'96, Ferreol-Chevalier'98, Belouchrani'98, Lee-Lewicki'99, deLathauwer'00, Pham-Cardoso'2000, Yeredor'2000, Sidiropoulos-Bro'00

General

General bibliography

- **Books on HOS, ICA, or Multi-Way:**

  Lacoume-Amblard-Comon'97

  Hyvarinen-Karhunen-Oja'01

  Smilde-Bro-Geladi'04

  Comon'07

- **Other related books:**

  Kagan-Linnik-Rao'73

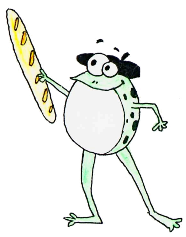
  McCullagh'87

  Nikias-Petropulu'93

  Haykin'2000

Contents

Part II: Tools



Algebraic tools

- PCA, SVD, Standardization
- Plane rotations, Jacobi sweeping

Statistical tools

- Mutual & Pairwise Independence
- Cumulants
- Mutual Information

Algebraic tools

Back to PCA

Definition

PCA is based on second order statistics

- Observed random variable $\boldsymbol{x}$ of dimension K. Then $\exists(\boldsymbol{U}, \boldsymbol{z})$:

$$\boldsymbol{x} = \boldsymbol{U}\boldsymbol{z},\ \boldsymbol{U} \text{ unitary}$$

  where *Principal Components* z_i are uncorrelated

  ith column $\boldsymbol{u}_i$ of $\boldsymbol{U}$ is called *ith PC Loading vector*

- Two possible calculations:
  - EVD of Covariance $\boldsymbol{R}_x$: $\boldsymbol{R}_x = \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^{\mathsf{H}}$
  - Sample estimate by SVD: $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathsf{H}}$

Algebraic tools

Standardization

Find a linear transform $\boldsymbol{L}$ such that vector $\tilde{\boldsymbol{x}} \stackrel{\text{def}}{=} \boldsymbol{L}\boldsymbol{x}$ has unit covariance. Many possibilities, including:

- PCA yields $\tilde{\boldsymbol{x}} = \boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathsf{H}}\boldsymbol{x}$
- Cholesky $\boldsymbol{R}_x = \boldsymbol{L}\,\boldsymbol{L}^{\mathsf{H}}$ yields $\tilde{\boldsymbol{x}} = \boldsymbol{L}^{-1}\boldsymbol{x}$

Remarks

- Infinitely many possibilities: $\boldsymbol{L}$ is as good as $\boldsymbol{L}\,\boldsymbol{Q}$, for any unitary $\boldsymbol{Q}$.
- If $\boldsymbol{R}_x$ not invertible, then $\boldsymbol{L}$ not invertible. One may use pseudo-inverse of $\boldsymbol{\Sigma}$ in PCA to compute $\boldsymbol{L}$.

Algebraic tools

Plane rotations

Application of a Givens rotation on both sides of a matrix allows to set a pair of zeros in a symmetric matrix;

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

Same result obtained:

- either by setting 0
- or by maximizing X's

Algebraic tools

Jacobi sweeping for PCA

Cyclic by rows/columns algorithm for a 4×4 real symmetric matrix

$$\begin{pmatrix} . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \end{pmatrix} \rightarrow \begin{pmatrix} X & 0 & x & x \\ 0 & X & x & x \\ x & x & . & . \\ x & x & . & . \end{pmatrix} \rightarrow \begin{pmatrix} X & x & 0 & x \\ x & . & x & . \\ 0 & x & X & x \\ x & . & x & . \end{pmatrix} \rightarrow \begin{pmatrix} X & x & x & 0 \\ x & . & . & x \\ x & . & . & x \\ 0 & x & x & X \end{pmatrix} \rightarrow$$

$$\begin{pmatrix} . & x & x & 0 \\ x & X & 0 & x \\ x & 0 & X & x \\ 0 & x & x & . \end{pmatrix} \rightarrow \begin{pmatrix} . & x & . & x \\ x & X & x & 0 \\ . & x & . & x \\ x & 0 & x & X \end{pmatrix} \rightarrow \begin{pmatrix} . & . & x & x \\ . & . & x & x \\ x & x & X & 0 \\ x & x & 0 & X \end{pmatrix}$$

X: maximized, x: minimized, 0: canceled, .: unchanged

Statistical independence

Definition

Components s_k of a K-dimensional r.v. $\boldsymbol{s}$ are *mutually independent*

$$\Updownarrow$$

The *joint* pdf equals the *product of marginal* pdf's:

$$p_{\boldsymbol{s}}(\boldsymbol{u}) = \prod_k p_{s_k}(u_k) \qquad (3)$$

Definition

Components s_k of $\boldsymbol{s}$ are *pairwise independent* $\Leftrightarrow$ Any pair of components (s_k, s_ℓ) are mutually independent.

Mutual vs Pairwise independence (1) $\ominus$

Example 3: Pairwise but not Mutual independence

- Bag containing 4 Bowls denoted {RB, YB, GB, RYB}:
  1 Red, 1 Yellow, 1 Green, 1 with the 3 colors.
- Equal drawing probabilities:
  $P(RB) = P(YB) = P(GB) = P(RYG) = 1/4$
- Event "R" $\stackrel{\text{def}}{=}$ draw a bowl containing Red $\Rightarrow$
  $P(R) = P(RB) + P(RYG) = 1/2$
- Then $P(R \cap Y) = P(RYG) = 1/4$
  equal to $P(R) * P(Y) \Rightarrow$ *Pairwise independent* Events
- But $P(R \cap Y \cap G) = P(RYG) = 1/4$
  not equal to $P(R) * P(Y) * P(G) = 1/8 \Rightarrow$
  Events are *not Mutually independent*

Statistical tools

Mutual vs Pairwise independence (2)

Example 4: Pairwise but not Mutual independence

- 3 mutually independent BPSK sources, $x_i \in \{-1, 1\}$, $1 \le i \le 3$
- Define $x_4 = x_1 x_2 x_3$. Then x_4 is also BPSK, *dependent on x_i*
- 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 $\Rightarrow$
  $p(x_2\, x_3 = b/a).p(x_1 = a) = \frac{1}{2} \cdot \frac{1}{2}$
- But x_k obviously not mutually independent, $1 \le k \le 4$
  In particular, $\text{Cum}\{x_1, x_2, x_3, x_4\} = 1 \neq 0$

Statistical tools

Mutual vs Pairwise independence (3)

Darmois's Theorem (1953)

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

$$Z_1 = \sum_{i=1}^{N} a_i\, x_i, \quad Z_2 = \sum_{i=1}^{N} b_i\, x_i$$

Then, if Z_1 and Z_2 are independent, those x_j for which $a_j b_j \neq 0$ are Gaussian.

Statistical tools

Mutual vs Pairwise independence (4)

Corollary

If $\boldsymbol{z} = \boldsymbol{C}\,\boldsymbol{s}$, where s_i are independent r.v., 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

Statistical tools

Characteristic functions

First

- Real Scalar: $\Phi_x(t) \stackrel{\text{def}}{=} \mathrm{E}\{e^{\jmath\, tx}\} = \int_u e^{\jmath\, tu}\, dF_x(u)$
- Real Multivariate: $\Phi_{\boldsymbol{x}}(\boldsymbol{t}) \stackrel{\text{def}}{=} \mathrm{E}\{e^{\jmath\, \boldsymbol{t}^{\mathsf{T}}\boldsymbol{x}}\} = \int_{\boldsymbol{u}} e^{\jmath\, \boldsymbol{t}^{\mathsf{T}}\boldsymbol{x}}\, dF_{\boldsymbol{x}}(\boldsymbol{u})$

Second

- $\Psi(\boldsymbol{t}) \stackrel{\text{def}}{=} \log \Phi(\boldsymbol{t})$
- Properties:
  - Always exists in the neighborhood of 0
  - Uniquely defined as long as $\Phi(\boldsymbol{t}) \neq 0$

Statistical tools

Cumulants (1)

- Moments:

$$\mu'_r \stackrel{\text{def}}{=} \mathrm{E}\{x^r\} = (-\jmath)^r \left.\frac{\partial^r \Phi(t)}{\partial t^r}\right|_{t=0} \tag{4}$$

- Cumulants:

$$\mathcal{C}_{x\,(r)} \stackrel{\text{def}}{=} \mathrm{Cum}\{\underbrace{x, \dots, x}_{r\ times}\} = (-\jmath)^r \left.\frac{\partial^r \Psi(t)}{\partial t^r}\right|_{t=0} \tag{5}$$

- Needs the existence of the expansion. Counter example: Cauchy

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

- Relationship between Moments and Cumulants obtained by expanding both sides in Taylor series:

$$\mathrm{Log}\,\Phi_x(t) = \Psi_x(t)$$

Statistical tools

Cumulants (2)

First Cumulants

- $\mathcal{C}_{(2)}$ is the variance:
- For zero-mean r.v.: $\mathcal{C}_{(3)} = \mu_{(3)}$, and $\mathcal{C}_{(4)} = \mu_{(4)} - 3\,\mu_{(2)}^2$
- Warning: it is not true that $\mathcal{C}_{(r)}$ is the moment of a variable $x - x_g$, x_g Gaussian
- Standardized cumulants:

$$\mathcal{K}_{(r)} = \mathrm{Cum}_{(r)}\left\{\frac{x - \mu'_{(1)}}{\sqrt{\mu_{(2)}}}\right\}$$

e.g. *Skewness* $\mathcal{K}_3$, and *Kurtosis* $\mathcal{K}_4$.

Examples of Cumulants (1)

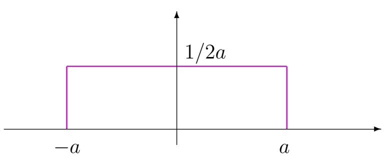
Example 5: Zero-mean Gaussian

- Moments
$$\mu_{(2r)} = \mu_{(2)}^r \frac{(2r)!}{r!\,2^r}$$
In particular: $\mu_{(4)} = 3\mu_{(2)}^2$, $\mu_{(6)} = 15\mu_{(2)}^3$.
- All Cumulants of order $r > 2$ are null

Examples of Cumulants (2)

Example 6: Uniform

- uniformly distributed in $[-a, +a]$ with probability $\frac{1}{2a}$
- Moments: $\mu_{2k} = \frac{a^{2k}}{2k+1}$
- 4th order Cumulant: $\mathcal{C}_4 = \frac{a^4}{5} - 3\,\frac{a^4}{9} = -2\,\frac{a^4}{15}$
- Kurtosis: $\mathcal{K}_4 = -\frac{6}{5}$.



Statistical tools

Examples of Cumulants (3)

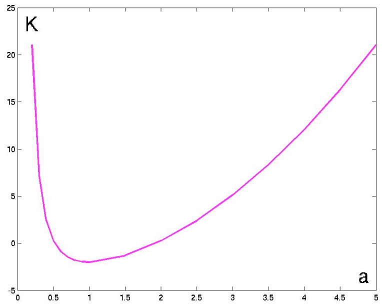
Example 7: Zero-mean standardized binary

- x takes two values $x_1 = -a$ and $x_2 = 1/a$ with probabilities $P_1 = \frac{1}{1+a^2}$, $P_2 = \frac{a^2}{1+a^2}$
- Skewness is $\mathcal{K}_{(3)} = \dfrac{1}{a} - a$
- Kurtosis is $\mathcal{K}_{(4)} = \dfrac{1}{a^2} + a^2 - 4$
- Extreme values

  Minimum Kurtosis

  for $a = 1$ (symmetric):

  $\mathcal{K}_{(4)} = -2$



Statistical tools

Multivariate cumulants

- Notation: $\mathcal{C}_{ij..\ell} \stackrel{\text{def}}{=} \text{Cum}\{X_i,\, X_j, ...X_\ell\}$
- First cumulants:

$$\begin{aligned} \mu'_i &= \mathcal{C}_i \\ \mu'_{ij} &= \mathcal{C}_{ij} + \mathcal{C}_i\mathcal{C}_j \\ \mu'_{ijk} &= \mathcal{C}_{ijk} + [3]\,\mathcal{C}_i\mathcal{C}_{jk} + \mathcal{C}_i\mathcal{C}_j\mathcal{C}_k \end{aligned}$$

  with $[n]$: Mccullagh's *bracket notation*.

- Next, for zero-mean variables:

$$\begin{aligned} \mu_{ijk\ell} &= \mathcal{C}_{ijk\ell} + [3]\,\mathcal{C}_{ij}\mathcal{C}_{k\ell} \\ \mu_{ijk\ell m} &= \mathcal{C}_{ijk\ell m} + [10]\,\mathcal{C}_{ij}\mathcal{C}_{k\ell m} \end{aligned}$$

- General formula of Leonov Shiryayev obtained by Taylor expansion of both sides of $\Psi(\boldsymbol{t}) = \log\Phi(\boldsymbol{t})$...

Statistical tools

Complex variables

Definition

Let $\boldsymbol{z} = \boldsymbol{x} + \jmath\,\boldsymbol{y}$. Then pdf $p_{\boldsymbol{z}}$ = joint pdf $p_{\boldsymbol{x},\boldsymbol{y}}$

Notation

- Characteristic function:

$$\Phi_{\boldsymbol{z}}(\boldsymbol{w}) = \mathrm{E}\{\exp[\jmath(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{u} + \boldsymbol{y}^{\mathsf{T}}\boldsymbol{v})]\} = \mathrm{E}\{\exp[\jmath\Re(\boldsymbol{z}^{\mathsf{H}}\boldsymbol{w})]\}$$

where $\boldsymbol{w} \stackrel{\text{def}}{=} \boldsymbol{u} + \boldsymbol{v}$.

- Generates Moments & Cumulants, e.g.:

Variance: $Var\{\boldsymbol{z}\}_{ij} = \mathcal{C}_{\boldsymbol{z}\,i}^{\;\;j}$

Higher orders: $\mathrm{Cum}\{z_i, \ldots, z_j, z_k^*, \ldots, z_\ell^*\} = \mathcal{C}_{\boldsymbol{z}\,i..j}^{\;\;k..\ell}$

where conjugated r.v. are labeled in superscript.

Statistical tools

Cumulant properties

- **Multi-linearity** (also enjoyed by moments):

$$\mathrm{Cum}\{\alpha X, Y, .., Z\} = \alpha\,\mathrm{Cum}\{X, Y, .., Z\} \quad (6)$$

$$\mathrm{Cum}\{X_1 + X_2, Y, .., Z\} = \mathrm{Cum}\{X_1, Y, .., Z\} + \mathrm{Cum}\{X_2, Y, .., Z\}$$

- **Cancellation**: If $\{X_i\}$ can be partitioned into 2 groups of independent r.v., then

$$\mathrm{Cum}\{X_1, X_2, .., X_r\} = 0 \quad (7)$$

- **Independence**: If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, then

$$\mathrm{Cum}\{X_1 + Y_1, X_2 + Y_2, .., X_r + Y_r\} = \mathrm{Cum}\{X_1, X_2, .., X_r\} + \mathrm{Cum}\{Y_1, Y_2, .., Y_r\}$$

- **Inequalities**, e.g.:

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

Statistical tools

Central Limit Theorem

Let N independent scalar r.v., $x(n), 1 \leq n \leq N$ each with finite rth order Cumulant, $\kappa_{(r)}(n)$.

Define:

$$\bar{\kappa}_{(r)} = \frac{1}{N}\sum_{n=1}^{N} \kappa_{(r)}(n) \text{ and } y = \frac{1}{\sqrt{N}}\sum_{n=1}^{N}(x(n) - \bar{\kappa}_{(1)}).$$

As $N \to \infty$, the pdf f_y tends to a Gaussian.

Proof:

$\mathcal{C}_{y\,(r)} = \frac{\bar{\kappa}_{(r)}}{N^{r/2-1}}$, $\forall r \geq 2$, tends to zero.

Statistical tools

Mutual Information (1)

- According to the definition of page 21, one should measure a divergence:

$$\delta\left(p_{\boldsymbol{x}}, \prod_{i=1}^{N} p_{x_i}\right)$$

- If the *Kullback divergence* is used:

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

then we get the *Mutual Information* as an independence measure:

$$I(p_{\boldsymbol{x}}) = \int p_{\boldsymbol{x}}(\boldsymbol{u}) \log \frac{p_{\boldsymbol{x}}(\boldsymbol{u})}{\prod_{i=1}^{N} p_{x_i}(u_i)}\, d\boldsymbol{u}. \qquad (8)$$

Statistical tools

Mutual Information (2)

Properties of the MI

- MI always positive
- Cancels if r.v. are mutually independent
- MI is invariant by scale change
- **Example 8: Gaussian case**

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

Statistical tools

Decomposition of the MI

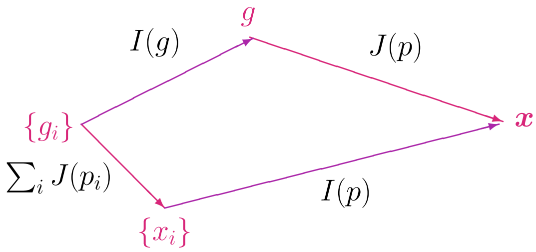
- Define the Negentropy as the divergence:

$$J(p_{\boldsymbol{x}}) = \delta(p_{\boldsymbol{x}}, g_{\boldsymbol{x}}) = \int p_{\boldsymbol{x}}(\boldsymbol{u})\, \log\frac{p_{\boldsymbol{x}}(\boldsymbol{u})}{g_{\boldsymbol{x}}(\boldsymbol{u})}\, d\boldsymbol{u}. \tag{9}$$

  Negentropy is invariant by invertible transforms

- Then MI can be decomposed into:

$$I(p_{\boldsymbol{x}}) = I(g_{\boldsymbol{x}}) + J(p_{\boldsymbol{x}}) - \sum_i J(p_{x_i}). \tag{10}$$



Statistical tools

Sample Measures of Statistical Independence ⊖

Independence at order r

- Definition:

  Components x_j of $\boldsymbol{x}$ are independent at order r if all *cross cumulants* of order r are null

- In other words: the *Cumulant tensor* $\mathcal{C}_{ij..\ell}$ is diagonal.

Example 9: Uncorrelated but not independent

$\boldsymbol{s}$ non Gaussian, s_i independent, then $\boldsymbol{x} = \boldsymbol{Q}\,\boldsymbol{s}$ has uncorrelated components *at order 2* if $\boldsymbol{Q}$ unitary $\longrightarrow$ cf. example slide 7.

Statistical tools

Edgeworth expansion (1)

Edgeworth expansion of a pdf

The pdf $p_{\boldsymbol{x}}(\boldsymbol{u})$ of a standardized r.v. $\boldsymbol{x}$ can be expanded about the Gaussian density $g_{\boldsymbol{x}}(\boldsymbol{u})$ of same mean and variance, in terms of a combination of Hermite polynomials, ordered by decreasing significance in the sense of the Central Limit Theorem (CLT).

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	$\kappa_4\kappa_5$	$\kappa_3^3\kappa_4$

From page 35, rth order Cumulants $\sim O(m^{1-r/2})$.

Statistical tools

Edgeworth expansion (2)

Edgeworth expansion of the MI

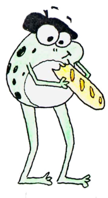
This yields for standardized random variables $\boldsymbol{x}$, after lengthy calculations:

$$I(p_{\boldsymbol{x}}) = J(p_{\boldsymbol{x}}) - \frac{1}{48}\sum_i 4\,\mathcal{C}_{iii}^{2} + \mathcal{C}_{iiii}^{2} + 7\,\mathcal{C}_{iii}^{4} - 6\,\mathcal{C}_{iii}^{2}\,\mathcal{C}_{iiii} + o(m^{-2}). \tag{11}$$

- If 3rd order $\neq 0$, then $I(p_{\boldsymbol{x}}) \approx J(p_{\boldsymbol{x}}) - \frac{1}{12}\sum_i \mathcal{C}_{iii}^{2}$
- If 3rd order ≈ 0, then $I(p_{\boldsymbol{x}}) \approx J(p_{\boldsymbol{x}}) - \frac{1}{48}\sum_i \mathcal{C}_{iiii}^{2}$

Contents

Part III: Separation of Independent Sources



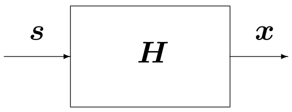
- Cumulant matching (direct approach: identification)
- Contrast Criteria (inverse approach: equalization):
- Numerical Algorithms: block/adaptive, joint/deflation

Cumulant Matching

Identification

Principle

- Estimate the mixture by solving the I/O Multi-linear equations: *Cumulant matching*



- Apply a separating filter based on the latter estimate

Cumulant Matching

Noiseless mixture of 2 sources

Example 10: 2×2 by Cumulant matching (cf. demo p.8)

- After standardization, the mixture takes the form

$$\boldsymbol{x} = \begin{pmatrix} \cos\alpha & -\sin\alpha\, e^{\jmath\varphi} \\ \sin\alpha\, e^{-\jmath\varphi} & \cos\alpha \end{pmatrix} \boldsymbol{s} \qquad (12)$$

- Denote $\gamma_{ij}^{k\ell} = \mathrm{Cum}\{x_i, x_j, x_k^*, x_\ell^*\}$ and $\kappa_i = \mathrm{Cum}\{s_i, s_i, s_i^*, s_i^*\}$. Then by *Multi-linearity*:

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

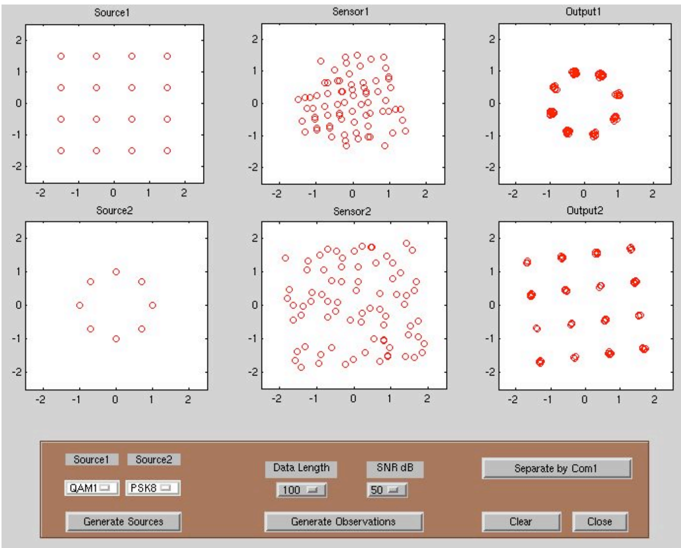
- Compact solution: $\frac{\gamma_{12}^{22} - \gamma_{11}^{12}}{\gamma_{12}^{12}} = -2\ \cot 2\alpha\, e^{\jmath\varphi}$

▷ demoICA2x2 ▷ Noiseless demo2C

Cumulant matching

Noiseless mixture of 2 sources

Example 11: Separation of 2 non Gaussian sources



Contrast criteria

Definition of a Contrast

Axiomatic definition

A *Contrast* optimization criterion Υ should enjoy 3 properties:

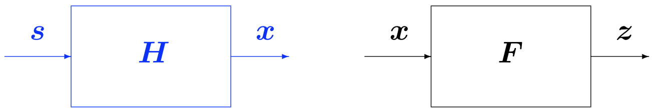
- *Invariance*: Υ should not change under the action of trivial filters (as defined in p.6)
- *Domination*: If sources are already separated, any filter should decrease (or leave unchanged) Υ
- *Discrimination*: The maximum achievable value should be reached only when sources are separated (i.e. maxima are related by trivial filters)

Contrast criteria

Mutual Information

$\Upsilon \stackrel{\text{def}}{=} -I(p_{\boldsymbol{z}})$ is a contrast

- Invariant by scale change and permutation
- Always negative
- Null if and only if components are independent



Contrast criteria

Maximum Likelihood

Given the source pdf's: $p_{\boldsymbol{s}}(\boldsymbol{u}) = \prod_i p_{s_i}(u_i)$, the ML approach consists of maximizing one of the criteria below

- **Noiseless case**

$$\mathcal{L} \stackrel{\text{def}}{=} p(\boldsymbol{x}|\boldsymbol{H}) = \frac{1}{|\det \boldsymbol{H}|} p_{\boldsymbol{s}}(\boldsymbol{H}^{-1}\boldsymbol{x})$$

- **Noisy case**

$$\mathcal{L} \stackrel{\text{def}}{=} p(\boldsymbol{x}, \boldsymbol{s}|\boldsymbol{H}) = g(\boldsymbol{x} - \boldsymbol{H}\,\boldsymbol{s})\, p_{\boldsymbol{s}}(\boldsymbol{s})$$

- And the *Joint MAP-ML* criterion for a joint estimation of sources:

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

Contrast criteria

Noiseless Maximum Likelihood (1)

- For an increasing number of independent observations, the average log-likelihood converges to

$$\frac{1}{T}\log p(\boldsymbol{x}_1 \dots \boldsymbol{x}_T|\boldsymbol{H}) \to \int \log p_{\boldsymbol{s}}(\boldsymbol{H}^{-1}\boldsymbol{u})\, p_{\boldsymbol{x}}(\boldsymbol{u})\, d\boldsymbol{u} + cst$$

which can be seen to be, by making the change $\boldsymbol{v} = \boldsymbol{H}^{-1}\boldsymbol{u}$:

$$\Upsilon_{ML} \stackrel{\text{def}}{=} -K(p_{\boldsymbol{z}}, p_{\boldsymbol{s}}) + cst \qquad (13)$$

☞ *pdf matching*

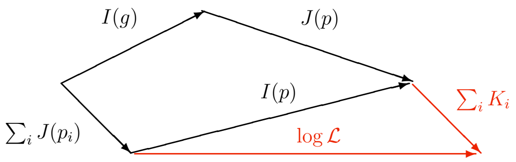
Contrast criteria

Noiseless Maximum Likelihood (2)

- Yet, since s_i are independent, it can be shown that

$$K(p_{\boldsymbol{z}}, p_{\boldsymbol{s}}) = \underbrace{K(p_{\boldsymbol{z}}, \prod_i p_{z_i})}_{MI} + \underbrace{\sum_i K(p_{z_i}, p_{s_i})}_{pdf\,deviation}$$

This allows to take into account the source pdf's, if they are known



- **But** ML is not adequate if source pdf's are unknown
  ⇒ just use contrast criteria, as MI

Contrast criteria

Contrasts of CoM(α, r)

When observations are standardized, and when only unitary transforms are considered, then the following are contrast functions:

- If at most 1 source has a null skewness:

$$\Upsilon_{2,3} = \sum_{p=1}^{P} (\kappa_{iii})^2, \quad \kappa_{iii} \stackrel{\text{def}}{=} \mathcal{C}_{\boldsymbol{z}\, iii}$$

- If at most 1 source has a null kurtosis:

$$\Upsilon_{2,4} = \sum_{p=1}^{P} (\kappa_{ii}^{ii})^2, \quad \kappa_{ii}^{ii} \stackrel{\text{def}}{=} \mathcal{C}_{\boldsymbol{z}\, ii}^{\;\; ii}$$

- If at most 1 source has a null standardized Cumulant of order $r > 2$, and for any $\alpha \geq 1$:

$$\Upsilon_{\alpha,r} = \sum_{p=1}^{P} |\kappa_{(r)}|^{\alpha}, \quad \kappa_{(r)} \stackrel{\text{def}}{=} \mathcal{C}_{\boldsymbol{z}\,(r)}$$

Contrast criteria

Contrast CoM$(1, 4)$

Example 12: Kurtosis-based contrast without squaring

- In particular, if all source kurtosis have the same sign, ε, one can avoid the absolute value:

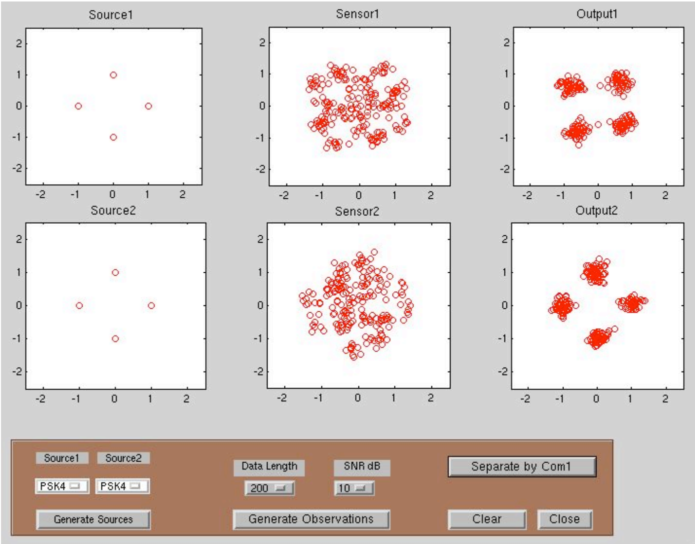
$$\Upsilon_{1,4} = \varepsilon \sum_{p=1}^{P} \kappa_{ii}^{ii}$$

▷ Noisy demo2C

Contrast criteria

Noisy mixture of 2 sources

Example 13: Separation of 2 non Gaussian sources by contrast maximization



Contrast criteria

JADE Contrast

Instead of minimizing all extra-diagonal terms:

$$\Theta - \Upsilon_{2,4} = \sum_{ijk\ell \neq iiii} |\mathcal{C}_{ij}^{k\ell}(\boldsymbol{Q}\,\tilde{\boldsymbol{x}})|^2$$

one minimizes

$$\Theta - \Upsilon_{Jade} = \sum_{ijk\ell \neq iik\ell} |\mathcal{C}_{ij}^{k\ell}(\boldsymbol{Q}\,\tilde{\boldsymbol{x}})|^2$$

which is equivalent to maximize $\Upsilon_{Jade} = \sum_{ik\ell} |\gamma_{i\ell}^{ik}|^2$.

Interest:

$$\Upsilon_{Jade} = \sum_{p=1}^{P^2} ||\mathbf{diag}(\boldsymbol{Q}^{\mathsf{H}}\,\boldsymbol{M}_r\,\boldsymbol{Q}||^2 \qquad (14)$$

is satisfied if the matrix set $\{\boldsymbol{M}_r\}$ forms an orthonormal basis.

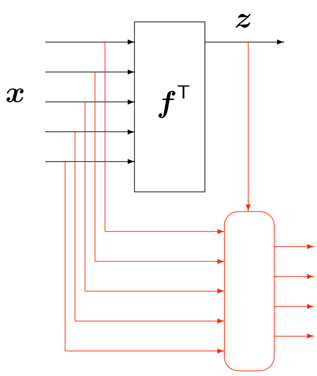
Algorithms

Block vs Adaptive

- Increase power of DSP
- Limitations of time-recursive Adaptive Algorithms
  - Convergence time of optimization algorithm
  - Convergence time of moment estimators
  - Local extrema harder to handle
- Coherence time sometimes limited
  (e.g. GSM: 900MHz, 190km/h, $T_c \approx 2ms \approx 300$ symbols)
- Well matched to block transmission (TDMA)
- Better exploitation of data
  (uniform weight, resistance to loss in synchro, time reversal)

Algorithms

Joint vs Deflation



Deflation:

- Advantage: reduced complexity at each stage
- Drawbacks: accumulation of regression errors, limitation of number of extracted sources

Algorithms

Deflation by Kurtosis Gradient Ascent

Adaptive Deflation by Kurtosis Maximization

After standardization, it is equivalent to maximize $4th$ order moment, $\mathrm{E}\{z^4\}$, which yields:

$$\Delta \boldsymbol{f} = \mu \, \nabla \mathcal{C}_{z\,(4)} = \mu \, \mathrm{E}\{\boldsymbol{x}\,(\boldsymbol{f}^{\mathsf{T}}\boldsymbol{x})^3\}$$

- After prewhitening, fixed step gradient on angles (Delfosse-Loubaton'95)
- "Locally optimal step" gradient on filter taps: FastICA (Hyvärinen'97)
- Globally optimal step gradient ascent (Comon'02)

Convergence: when $\boldsymbol{f}$ and $\nabla \mathcal{C}_{z\,(4)}$ collinear (and *not* when gradient is null, because of constraint $||\boldsymbol{f}|| = 1$).

Algorithms

Jacobi Sweeping

Joint Block Algorithm: Sweeping a $3 \times 3 \times 3$ tensor

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

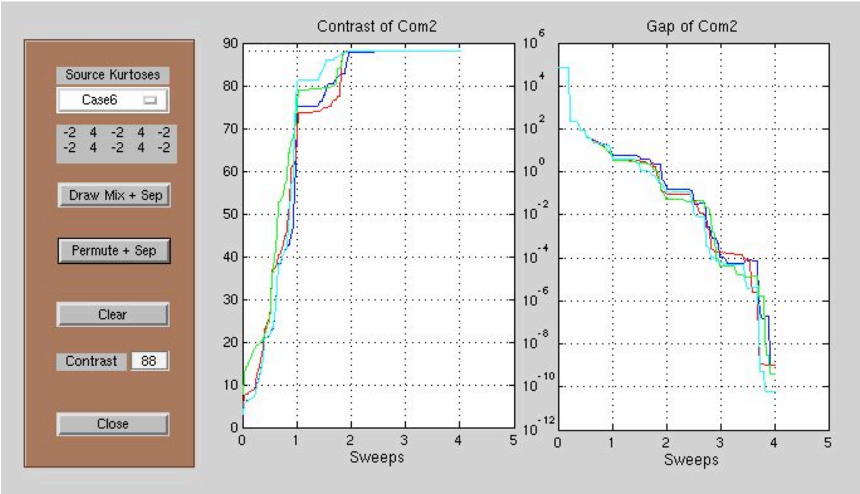
$$\left.\begin{array}{rl} X: & \text{maximized} \\ x: & \text{minimized} \\ .: & \text{unchanged} \end{array}\right\} \text{by the last Givens rotation}$$

▷ demo10R

Algorithms

Influence on Sweeping oder

Example 14: The order does not affect the limit, despite the presence of local maxima



Algorithms

Tensors as Linear Operators

- Linear Operator Ω acting on square matrices:

$$\boldsymbol{M} \longrightarrow \Omega(\boldsymbol{M})_{ij} = \sum_{k\ell} \mathcal{C}_{ik}^{j\ell} M_{k\ell}$$

admits eigen-matrices $\boldsymbol{N}_r$, $1 \leq r \leq P^2$.

- In the absence of noise, P nonzero eigenvalues
- In practice, retain P dominant eigen-matrices $\Rightarrow$ (i) reduced complexity P^2, and (ii) noise reduction
- **A Joint Block Algorithm:** JADE
  - Maximize $\Upsilon_{\alpha,Jade} \stackrel{\text{def}}{=} \sum_r ||\lambda_r^\alpha \mathbf{diag}(\boldsymbol{U}^{\mathsf{H}} \boldsymbol{N}_r \boldsymbol{U})||^2$
    "Joint Approximate Diagonalization of Eigenmatrices"
  - Sweep the pairs $\longrightarrow$ again a quadratic form

Contents

Part IV: Under-Determined Mixtures

What is specific:

- No linear inverse exists (thus no contrast)
- Prior standardization of poor usefulness

Two families of approaches:

- From Cumulant tensor
- From Data tensor

UDM from Cumulant tensor

Canonical Decomposition

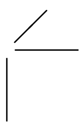
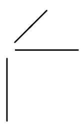
Cumulant Tensor Matching (example at order 3):

- Model + Multi-linearity yields:

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

- Canonical Tensor Decomposition (CanD):

$$\boldsymbol{T} = \sum_{p=1}^{rank(\boldsymbol{T})} \kappa_p\, \boldsymbol{h}(p) \circ \boldsymbol{h}(p) \circ \boldsymbol{h}(p) + \boldsymbol{E} \qquad (15)$$

$\boldsymbol{T} = \kappa_1$  $+ \cdots + \kappa_P$ 

- In practice, often minimize the matching error $\Psi \stackrel{\text{def}}{=} ||\boldsymbol{E}||^2$

Tensor Rank (1)

- Generic/Typical rank ω of symmetric tensors of order d, generally larger than dimension K:

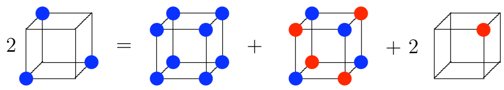
ω	K	**2**	**3**	**4**	**5**	**6**	**7**	**8**
d	**3**	2	4	5	8	10	12	15
	4	3	6	10	15	22	30	42

- CanD often not unique (in red: infinitely many solutions)

Tensor rank (2)

- Maximal rank: generally larger than generic rank

Example 15: order 3, dimension 2, but rank 3



blue bullets = 1, red bullets = −1.

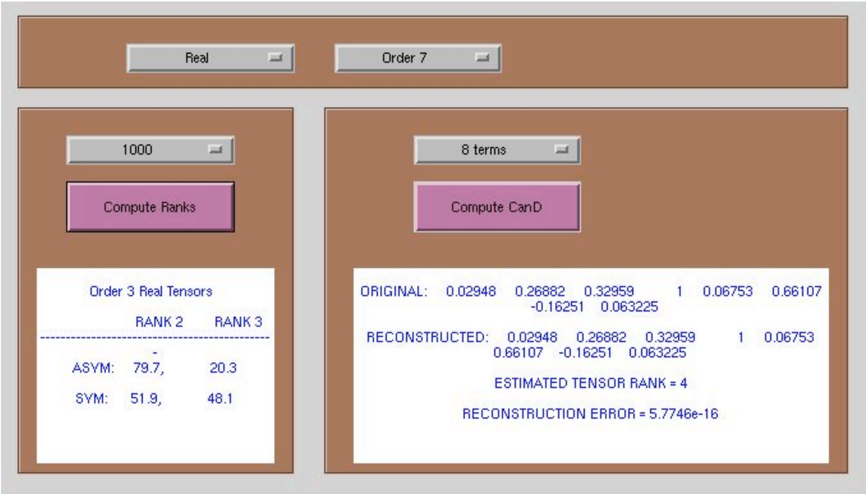
- In dimension 2, CanD entirely computable thanks to Sylvester's theorem on polynomials
- Very hard in higher dimensions

▷ demo BinaryTensors

UDM from Cumulant tensor

CanD of 2-dim tensors

Example 16: Rank obtained for dth order symmetric tensors of dim 2



UDM from Cumulant tensor

Tensor rank (3)

- Real tensors may not have same rank if immersed in complex field.

Example 17: 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 $\mathbb{C}$:

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

UDM from Cumulant tensor

Source extraction

Example 18: 3 BPSK sources and 2 sensors

- s_1, $_2$, s_3: $\in \{-1, 1\}$, mutually independent
- Actual observations: $\boldsymbol{x} = [x_1,\, x_2]^{\mathsf{T}}$
- Build virtual observations: $\boldsymbol{z} = [x_1^3,\, x_1^2 x_2,\, x_1 x_2^2,\, x_2^3]^{\mathsf{T}}$
- Then 6-dimensional augmented observation:

$$\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{z} \end{pmatrix} = \begin{bmatrix} \boldsymbol{H} & \boldsymbol{0} \\ \multicolumn{2}{c}{\boldsymbol{B}} \end{bmatrix} \cdot \begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_1 s_2 s_3 \end{pmatrix}$$

with one virtual source $s_4 \stackrel{\text{def}}{=} s_1 s_2 s_3$, pairwise independent of s_i

UDM form Data tensor

LS Criterion

Data arranged in a order-d tensor (d-way array)

- CanD in the case of $d = 3$:

$$\boldsymbol{T} = \sum_{p=1}^{\omega} \kappa_p\, \boldsymbol{a}(p) \circ \boldsymbol{b}(p) \circ \boldsymbol{c}(p) \tag{16}$$

- Now error Ψ is quadratic in each $\boldsymbol{a}(p)$, if all $\boldsymbol{b}(p)$ and $\boldsymbol{c}(p)$ fixed
- Other useful writings:

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

$$\Psi = ||\boldsymbol{T}^{(1)} - \boldsymbol{A}(\boldsymbol{C} \odot \boldsymbol{B})^{\mathsf{T}}||^2 \tag{17}$$

Minimum of Ψ w.r.t. $\boldsymbol{A}$ can be obtained by SVD.

Idem for $\boldsymbol{B}$, $\boldsymbol{C}$.

UDM form Data tensor

Alternate Least Squares (ALS)

The PARAFAC algorithm computes in turn $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$: Alternating Least Squares (ALS)

- Very slow convergence
- Need for a sufficient condition of uniqueness:

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

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

☞In symmetric case, one needs at least that $2\omega \leq 3K - 2$

☞Can be extended to order d: $2\omega \leq dK - d + 1$

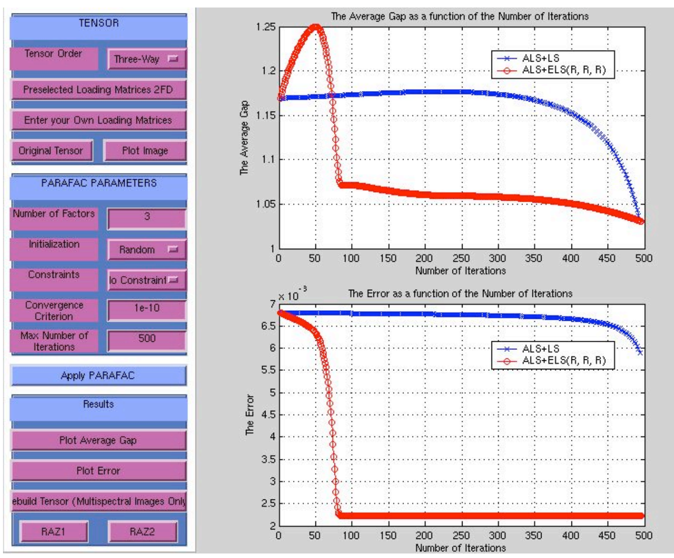
- Need for *diversity*: matrix slices must be "sufficiently different"

▷ demo Parafac

UDM from Data tensor

Parafac ALS algorithm

Example 19: Two accelerated versions: Bro'98 and Rajih-Comon'05



UDM form Data tensor

Kruskal rank

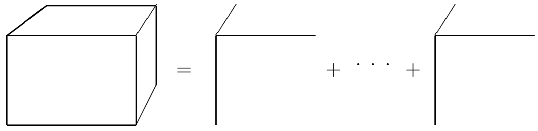
- **Column rank** of a matrix
  $rank(\boldsymbol{A}) = r$ iff there is *at least one* subset of r lin. independent columns, and this *fails for any subset* of $r+1$ columns.
- **Kruskal rank** of a matrix
  $K - rank(\boldsymbol{A}) = k$ iff *every* subset of k columns is lin. independent, and this *fails for at least one subset* of $k+1$ columns.
- Property: $k(\boldsymbol{A}) \leq rank(\boldsymbol{A}) \leq dim(\boldsymbol{A})$

UDM form Data tensor

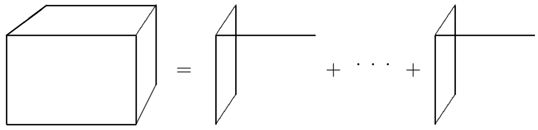
Data vs Cumulant Tensors

Multi-linear vs Linear Blind Model fitting

- CanD, if diversity among loading vectors allows to build a data tensor:



- ICA, if little diversity imposes a 2-way equivalent data matrix



Contents

Part V: Beyond this Tutorial

- Some unaddressed problems
- Tensor properties
- False beliefs

Beyond this Tutorial

Some Unaddressed Problems

- Reduction of tensor sizes: HOSVD/Tucker3 model fitting
- Simultaneous Tensor Diagonalization (STD)
- Performance indices
- Nonstationary sources, Discrete sources
- Convolutive mixtures
- Semi-Blind approaches
- Unexpected topological properties of tensor spaces

Beyond this Tutorial

Unexpected topological properties

- The variety of rank-1 matrices or tensors is closed
- The variety of matrices of rank $\leq k$ is closed
- The set of tensors of rank $\leq k$ **is not** closed; e.g.:

  $\exists$ sequence $\boldsymbol{T}_n$ of rank-3 tensors $\Rightarrow$ rank 4 !

Beyond this Tutorial

False Beliefs (1)

1. BSS always requires High-Order Statistics (HOS)

$\longrightarrow$ *Second-order can (rarely) suffice*

2. Sources must be statistically independent

$\longrightarrow$ *Correlated sources can be sometimes separated (e.g. Discrete/CM sources, Pairwise cumulants...)*

3. HOS are always required when sources are *i.i.d.*

$\longrightarrow$ *Second-order BSS algorithms exist*

4. There should be at least as many sensors as sources: $K \geq P$ (sufficient diversity)

$\longrightarrow$ *Underdetermined mixtures can be identified*

Beyond this Tutorial

False Beliefs (2)

5. Perfect source extraction is impossible if $K < P$

$\longrightarrow$ *Discrete sources can be perfectly extracted from underdetermined mixtures (insufficient diversity)*

6. Conditions of application of Parafac are mild

$\longrightarrow$ *except when one dimension* $= 2$, *the typical rank always exceeds the Parafac bound for uniqueness*

7. Approximate a tensor by another of lower rank is as easy as for matrices

$\longrightarrow$ *beside for rank 1, there is a lack of closeness*

8. The Constant Modulus (CM) property is the best way to handle PSK sources

$\longrightarrow$ *The whole alphabet can be taken into account in order to define a contrast function*