A Very Short Introduction to Blind Source Separation

a.k.a. How You Will Definitively Enjoy Differently a Cocktail Party

Matthieu Puigt

Foundation for Research and Technology – Hellas
Institute of Computer Science
mpuigt@forth.ics.gr
http://www.ics.forth.gr/~mpuigt

April 12 / May 3, 2011

Let's talk about linear systems

All of you know how to solve this kind of systems:

$$\begin{cases} 2 \cdot s_1 + 3 \cdot s_2 & = & 5 \\ 3 \cdot s_1 - 2 \cdot s_2 & = & 1 \end{cases} \tag{1}$$

If we resp. define A, $\underline{s}$, and $\underline{x}$ the matrix and the vectors:

$$A = \begin{bmatrix} 2 & 3 \\ 3 & -2 \end{bmatrix}, \underline{s} = [s_1, s_2]^T, \text{ and } \underline{x} = [5, 1]^T$$

Eq. (1) begins

$$\underline{x} = A \cdot \underline{s}$$

and the solution reads:

$$\underline{s} = A^{-1} \cdot \underline{x} = [1, 1]^T$$

Let's talk about linear systems

All of you know how to solve this kind of systems:

$$\left\{ \begin{array}{rcl} {\color{red}?} \cdot s_1 + {\color{red}?} \cdot s_2 & = & 5 \\ {\color{red}?} \cdot s_1 + {\color{red}?} \cdot s_2 & = & 1 \end{array} \right. \tag{1}$$

If we resp. define A, $\underline{s}$, and $\underline{x}$ the matrix and the vectors:

$$A = \begin{bmatrix} {\color{red}?} & {\color{red}?} \\ {\color{red}?} & {\color{red}?} \end{bmatrix}, \underline{s} = [s_1, s_2]^T, \text{ and } \underline{x} = [5, 1]^T$$

Eq. (1) begins

$$\underline{x} = A \cdot \underline{s}$$

and the solution reads:

$$\underline{s} = A^{-1} \cdot \underline{x} = {\color{red}?}$$

Let's talk about linear systems

All of you know how to solve this kind of systems:

$$\begin{cases} a_{11} \cdot s_1 + a_{12} \cdot s_2 & = & 5 \\ a_{21} \cdot s_1 + a_{22} \cdot s_2 & = & 1 \end{cases} \tag{1}$$

If we resp. define A, $\underline{s}$, and $\underline{x}$ the matrix and the vectors:

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \underline{s} = [s_1, s_2]^T, \text{ and } \underline{x} = [5, 1]^T$$

Eq. (1) begins

$$\underline{x} = A \cdot \underline{s}$$

and the solution reads:

$$\underline{s} = A^{-1} \cdot \underline{x} = \mathbf{?}$$

How can we solve this kind of problem???

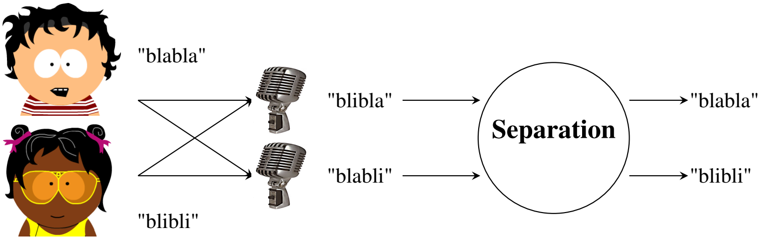
This problem is called **Blind Source Separation**.

Blind Source Separation problem

- N unknown sources s_j.
- One unknown operator A.
- P observed signals x_i with the global relation

$$\underline{x} = A\left(\underline{s}\right).$$

Goal: Estimating the vector $\underline{s}$, up to some indeterminacies.



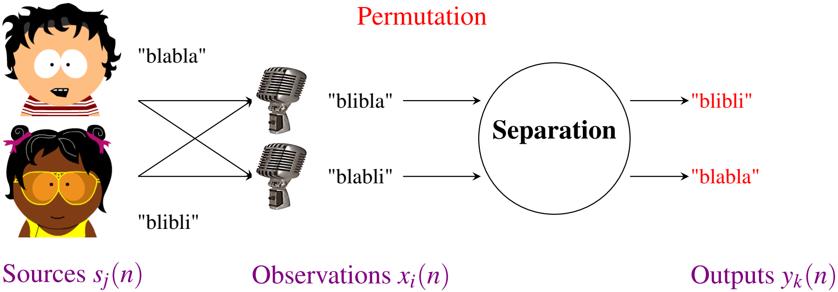
Sources $s_j(n)$ Observations $x_i(n)$ Outputs $y_k(n)$

Blind Source Separation problem

- N unknown sources s_j.
- One unknown operator A.
- P observed signals x_i with the global relation

$$\underline{x} = A\left(\underline{s}\right).$$

Goal: Estimating the vector $\underline{s}$, up to some indeterminacies.

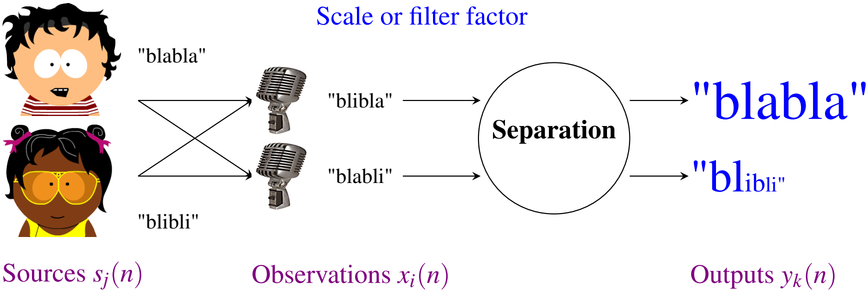


Blind Source Separation problem

- N unknown sources s_j.
- One unknown operator A.
- P observed signals x_i with the global relation

$$\underline{x} = A\left(\underline{s}\right).$$

Goal: Estimating the vector $\underline{s}$, up to some indeterminacies.

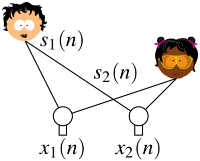


Most of the approaches process linear mixtures which are divided in three categories:

1. Linear instantaneous (LI) mixtures: $x_i(n) = \sum_{j=1}^{N} a_{ij}\, s_j(n)$ (Purpose of this lecture)

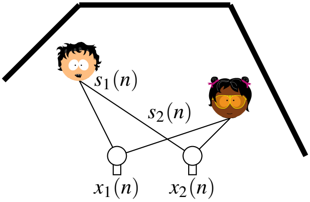
Most of the approaches process linear mixtures which are divided in three categories:

1. Linear instantaneous (LI) mixtures: $x_i(n) = \sum_{j=1}^{N} a_{ij}\, s_j(n)$ (Purpose of this lecture)
2. Attenuated and delayed (AD) mixtures: $x_i(n) = \sum_{j=1}^{N} a_{ij}\, s_j(n - n_{ij})$



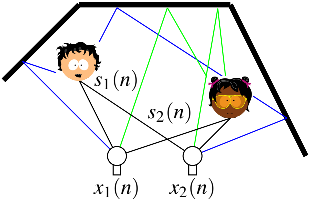
Most of the approaches process linear mixtures which are divided in three categories:

1. Linear instantaneous (LI) mixtures: $x_i(n) = \sum_{j=1}^{N} a_{ij}\, s_j(n)$ (Purpose of this lecture)
2. Attenuated and delayed (AD) mixtures: $x_i(n) = \sum_{j=1}^{N} a_{ij}\, s_j(n - n_{ij})$
3. Convolutive mixtures:
   $x_i(n) = \sum_{j=1}^{N} \sum_{k=-\infty}^{+\infty} a_{ijk}\, s_j(n - n_{ijk}) = \sum_{j=1}^{N} a_{ij}(n) * s_j(n)$



Most of the approaches process linear mixtures which are divided in three categories:

1. Linear instantaneous (LI) mixtures: $x_i(n) = \sum_{j=1}^{N} a_{ij}\, s_j(n)$ (Purpose of this lecture)
2. Attenuated and delayed (AD) mixtures: $x_i(n) = \sum_{j=1}^{N} a_{ij}\, s_j(n - n_{ij})$
3. Convolutive mixtures:
$x_i(n) = \sum_{j=1}^{N} \sum_{k=-\infty}^{+\infty} a_{ijk}\, s_j(n - n_{ijk}) = \sum_{j=1}^{N} a_{ij}(n) * s_j(n)$



Let's go back to our previous problem

A kind of magic?

- Here, the operator is a simple matrix whose coefficients are unknown.

$$\begin{cases} a_{11} \cdot s_1 + a_{12} \cdot s_2 & = & 5 \\ a_{21} \cdot s_1 + a_{22} \cdot s_2 & = & 1 \end{cases}$$

- In Signal Processing, we do not have the unique above system of equation but a **series** of such systems (due to **samples**)

We thus use the intrinsic properties of source signals to achieve the separation (assumptions)

Let's go back to our previous problem

A kind of magic?

- Here, the operator is a simple matrix whose coefficients are unknown.

$$\begin{cases} a_{11} \cdot s_1 + a_{12} \cdot s_2 & = & 0 \\ a_{21} \cdot s_1 + a_{22} \cdot s_2 & = & .24 \end{cases}$$

- In Signal Processing, we do not have the unique above system of equation but a **series** of such systems (due to **samples**)

We thus use the intrinsic properties of source signals to achieve the separation (assumptions)

Let's go back to our previous problem

A kind of magic?

- Here, the operator is a simple matrix whose coefficients are unknown.

$$\begin{cases} a_{11} \cdot s_1 + a_{12} \cdot s_2 & = & 4 \\ a_{21} \cdot s_1 + a_{22} \cdot s_2 & = & -2 \end{cases}$$

- In Signal Processing, we do not have the unique above system of equation but a **series** of such systems (due to **samples**)

We thus use the intrinsic properties of source signals to achieve the separation (assumptions)

Let's go back to our previous problem

A kind of magic?

- Here, the operator is a simple matrix whose coefficients are unknown.

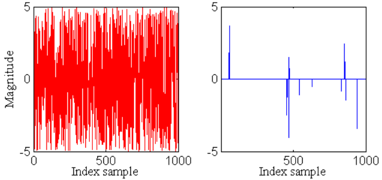
$$\begin{cases} a_{11} \cdot s_1(n) + a_{12} \cdot s_2(n) & = & x_1(n) \\ a_{21} \cdot s_1(n) + a_{22} \cdot s_2(n) & = & x_2(n) \end{cases}$$

- In Signal Processing, we do not have the unique above system of equation but a **series** of such systems (due to **samples**)

We thus use the intrinsic properties of source signals to achieve the separation (assumptions)

Three main families of methods:

1. **Independent Component Analysis (ICA):** Sources are statistically independent, stationary and at most one of them is Gaussian (in their basic versions).
2. **Sparse Component Analysis (SCA):** Sparse sources (i.e. most of the samples are null (or close to zero)).
3. **Non-negative Matrix Factorization (NMF):** Both sources et mixtures are positive, with possibly sparsity constraints.



A bit of history (1)

- BSS problem formulated around 1982, by Hans, Hérault, and Jutten for a biomedical problem and first papers in the mid of the 80's
- Great interest from the community, mainly in France and later in Europe and in Japan, and then in the USA
  - Several special sessions in international conferences (e.g. GRETSI'93, NOLTA'95, etc)
  - First workshop in 1999, in Aussois, France. One conference each 18 months (see `http://research.ics.tkk.fi/ica/links.shtml`) and next one in 2012 in Tel Aviv, Israel
  - *"In June 2009, 22000 scientific papers are recorded by Google Scholar"* (Comon and Jutten, 2010)
  - People with different backgrounds: signal processing, statistics, neural networks, and later machine learning
- Initially, BSS addressed for LI mixtures but
  - convolutive mixtures in the mid of the 90's
  - nonlinear mixtures at the end of the 90's
- Until the end of the 90's, BSS $\simeq$ ICA
  - First NMF methods in the mid of the 90's but famous contribution in 1999
  - First SCA approaches around 2000 but massive interest since

A bit of history (2)

- BSS on the web:
  - Mailing list in ICA Central: http://www.tsi.enst.fr/icacentral/
  - Many softwares available in ICA Central, ICALab (http://www.bsp.brain.riken.go.jp/ICALAB/), NMFLab (www.bsp.brain.riken.go.jp/ICALAB/nmflab.html), etc.
  - International challenges:
    1. 2006 Speech Separation Challenge (http://staffwww.dcs.shef.ac.uk/people/M.Cooke/SpeechSeparationChallenge.htm)
    2. 2007 MLSP Competition (http://mlsp2007.conwiz.dk/index.php@id=43.html)
    3. Signal Separation Evaluation Campaigns in 2007, 2008, 2010, and 2011 (http://sisec.wiki.irisa.fr/tiki-index.php)
    4. 2011 Pascal CHIME Speech separation and recognition (http://www.dcs.shef.ac.uk/spandh/chime/challenge.html)

A “generic” problem

Many applications: biomedical, audio processing and audio coding, telecommunications, astrophysics, image classification, underwater acoustics, finance, etc.

Content of the lecture

1. Sparse Component Analysis
2. Independent Component Analysis
3. Non-negative Matrix Factorization?

Some good documents

- P. Comon and C. Jutten: *Handbook of Blind Source Separation. Independent component analysis and applications*. Academic Press (2010)
- A. Hyvärinen, J. Karhunen, and E. Oja: *Independent Component Analysis*. Wiley-Interscience, New York (2001)
- S. Makino, T.W. Lee, and H. Sawada: *Blind Speech Separation*. Signals and Communication Technology, Springer (2007)
- Wikipedia: `http://en.wikipedia.org/wiki/Blind_signal_separation`
- Many online tutorials...

Part I

Sparse Component Analysis

Let's go back to our previous problem

We said we have a series of systems of equations. Let's denote $x_i(n)$ and $s_j(n)$ $(1 \leq i \leq j \leq 2)$ the values that take both source and observation signals.

$$\begin{cases} a_{11} \cdot s_1(n) + a_{12} \cdot s_2(n) = x_1(n) \\ a_{21} \cdot s_1(n) + a_{22} \cdot s_2(n) = x_2(n) \end{cases}$$

SCA methods main idea

- Sources are sparse, i.e. often zero.
- We assume that $a_{11} \neq 0$ and that $a_{12} \neq 0$
- We thus have a lot of chances that for one given index n_0, one source (say $s_1(n_0)$) is the only **active** source. In this case, the system is much simpler.

Let's go back to our previous problem

We said we have a series of systems of equations. Let's denote $x_i(n)$ and $s_j(n)$ $(1 \leq i \leq j \leq 2)$ the values that take both source and observation signals.

$$\begin{cases} a_{11} \cdot s_1(n_0) & = & x_1(n_0) \\ a_{21} \cdot s_1(n_0) & = & x_2(n_0) \end{cases}$$

SCA methods main idea

- Sources are sparse, i.e. often zero.
- We assume that $a_{11} \neq 0$ and that $a_{12} \neq 0$
- We thus have a lot of chances that for one given index n_0, one source (say $s_1(n_0)$) is the only **active** source. In this case, the system is much simpler.

Let's go back to our previous problem

We said we have a series of systems of equations. Let's denote $x_i(n)$ and $s_j(n)$ $(1 \leq i \leq j \leq 2)$ the values that take both source and observation signals.

$$\left\{ \begin{array}{lcl} a_{11} \cdot s_1(n_0) & = & x_1(n_0) \\ a_{21} \cdot s_1(n_0) & = & x_2(n_0) \end{array} \right.$$

SCA methods main idea

- Sources are sparse, i.e. often zero.
- We assume that $a_{11} \neq 0$ and that $a_{12} \neq 0$
- We thus have a lot of chances that for one given index n_0, one source (say $s_1(n_0)$) is the only **active** source. In this case, the system is much simpler.
- If we compute the ratio $\frac{x_2(n_0)}{x_1(n_0)}$, we obtain: $\frac{x_2(n_0)}{x_1(n_0)} = \frac{a_{21} \cdot s_1(n_0)}{a_{11} \cdot s_1(n_0)} = \frac{a_{21}}{a_{11}}$
- Instead of $[a_{11}, a_{21}]^T$, we thus can estimate $\left[1, \frac{a_{21}}{a_{11}}\right]^T$
- Let us see why!

- Imagine now that, for each source, we have (at least) one sample (**single-source** samples) for which only one source is active:

$$\begin{cases} a_{11} \cdot s_1(n\ ) + a_{12} \cdot s_2(n\ ) & = & x_1(n\ ) \\ a_{21} \cdot s_1(n\ ) + a_{22} \cdot s_2(n\ ) & = & x_2(n\ ) \end{cases} \tag{2}$$

- Imagine now that, for each source, we have (at least) one sample (**single-source** samples) for which only one source is active:

$$\left\{\begin{array}{lcl} a_{11} \cdot s_1(n_0) & = & x_1(n_0) \\ a_{21} \cdot s_1(n_0) & = & x_2(n_0) \end{array}\right. \tag{2}$$

- Imagine now that, for each source, we have (at least) one sample (**single-source** samples) for which only one source is active:

$$\begin{cases} a_{12} \cdot s_2(n_1) = x_1(n_1) \\ a_{22} \cdot s_2(n_1) = x_2(n_1) \end{cases} \tag{2}$$

- Imagine now that, for each source, we have (at least) one sample (**single-source** samples) for which only one source is active:

$$\begin{cases} a_{11} \cdot s_1(n\ ) + a_{12} \cdot s_2(n\ ) & = & x_1(n\ ) \\ a_{21} \cdot s_1(n\ ) + a_{22} \cdot s_2(n\ ) & = & x_2(n\ ) \end{cases} \quad (2)$$

- Ratio $\frac{x_2(n)}{x_1(n)}$ for samples n_0 and n_1 $\Rightarrow$ scaled version of A, denoted B:

$$B = \begin{bmatrix} 1 & 1 \\ \frac{a_{21}}{a_{11}} & \frac{a_{22}}{a_{12}} \end{bmatrix} \text{ or } B = \begin{bmatrix} 1 & 1 \\ \frac{a_{22}}{a_{12}} & \frac{a_{21}}{a_{11}} \end{bmatrix}$$

- If we express Eq. (2) in matrix form with respect to B, we read:

$$\underline{x}(n) = B \cdot \begin{bmatrix} a_{11} \cdot s_1(n) \\ a_{12} \cdot s_2(n) \end{bmatrix} \text{ or } \underline{x}(n) = B \cdot \begin{bmatrix} a_{12} \cdot s_2(n) \\ a_{11} \cdot s_1(n) \end{bmatrix}$$

- and by left-multiplying by B^{-1}:

$$\underline{y}(n) = B^{-1} \cdot \underline{x}(n) = B^{-1} \cdot B \cdot \begin{bmatrix} a_{11} \cdot s_1(n) \\ a_{12} \cdot s_2(n) \end{bmatrix} = \begin{bmatrix} a_{11} \cdot s_1(n) \\ a_{12} \cdot s_2(n) \end{bmatrix}$$

$$\text{or } \underline{y}(n) = B^{-1} \cdot \underline{x}(n) = B^{-1} \cdot B \cdot \begin{bmatrix} a_{12} \cdot s_2(n) \\ a_{11} \cdot s_1(n) \end{bmatrix} = \begin{bmatrix} a_{12} \cdot s_2(n) \\ a_{11} \cdot s_1(n) \end{bmatrix}$$

How to find single-source samples?

Different assumptions

- Strong assumption: sources **W-disjoint orthogonal** (WDO), i.e. in each sample, only one source is active.
- Weak assumption: several sources active in the same samples, **except for some tiny zones** (to find) where only one source occurs.
- Hybrid assumption: sources WDO but single-source confidence measure to accurately estimate the mixing parameters.

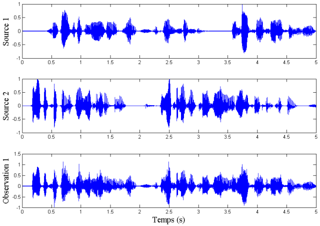
How to find single-source samples?

Different assumptions

- Strong assumption: sources **W-disjoint orthogonal** (WDO), i.e. in each sample, only one source is active.
- Weak assumption: several sources active in the same samples, **except for some tiny zones** (to find) where only one source occurs.
- Hybrid assumption: sources WDO but single-source confidence measure to accurately estimate the mixing parameters.

TEMPROM (Abrard *et al.*, 2001)

- TEMPROM: **TEMP**oral **R**atio **O**f **M**ixtures
- Main steps:
  1. Detection stage: finding single-source zones
  2. Identification stage: estimating B
  3. Reconstruction stage: recovering the sources



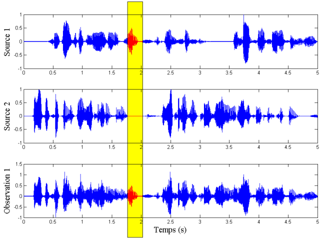
How to find single-source samples?

Different assumptions

- Strong assumption: sources **W-disjoint orthogonal** (WDO), i.e. in each sample, only one source is active.
- Weak assumption: several sources active in the same samples, **except for some tiny zones** (to find) where only one source occurs.
- Hybrid assumption: sources WDO but single-source confidence measure to accurately estimate the mixing parameters.

TEMPROM (Abrard *et al.*, 2001)

- TEMPROM: **TEMP**oral **R**atio **O**f **M**ixtures
- Main steps:
  1. Detection stage: finding single-source zones
  2. Identification stage: estimating B
  3. Reconstruction stage: recovering the sources



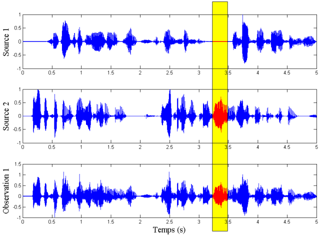
How to find single-source samples?

Different assumptions

- Strong assumption: sources **W-disjoint orthogonal** (WDO), i.e. in each sample, only one source is active.
- Weak assumption: several sources active in the same samples, **except for some tiny zones** (to find) where only one source occurs.
- Hybrid assumption: sources WDO but single-source confidence measure to accurately estimate the mixing parameters.

TEMPROM (Abrard *et al.*, 2001)

- TEMPROM: **TEMP**oral **R**atio **O**f **M**ixtures
- Main steps:
  1. Detection stage: finding single-source zones
  2. Identification stage: estimating B
  3. Reconstruction stage: recovering the sources



TEMPROM detection stage

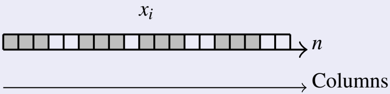
- Let's go back to our problem with 2 sources and 2 observations.
- Imagine that in one zone $T = \{n_1, \ldots, n_M\}$, only one source, say s_1 is active.
- According to what we saw, the ratio $\frac{x_2(n)}{x_1(n)}$ on this zone is equal to $\frac{a_{21}}{a_{11}}$ and is thus constant.
- On the contrary, if both sources are active, this ratio varies.
- The **variance** of this ratio over time zones is thus a single-source confidence measure: lowest values correspond to single-source zones!

Steps of detection stage

1. Cutting the signals in small temporal zones
2. Computing the variance over these zones of the ratio $\frac{x_2(n)}{x_1(n)}$
3. Ordering the zones according to increasing variance of the ratio

TEMPROM identification stage

1. Successively considering the zones in the above sorted list
2. Estimating a new column (average over the considered zone of the ratio $\frac{x_2}{x_1}$)
3. Keeping it if its distance wrt previously found ones is "sufficiently high"
4. Stoping when all the columns of B are estimated



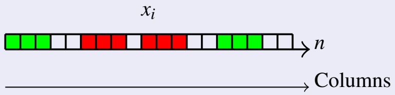
Problem

Finding such zones is hard with:

- continuous speech (non-civilized talk where everyone speaks at the same time)
- short pieces of music

TEMPROM identification stage

1. Successively considering the zones in the above sorted list
2. Estimating a new column (average over the considered zone of the ratio $\frac{x_2}{x_1}$)
3. Keeping it if its distance wrt previously found ones is "sufficiently high"
4. Stoping when all the columns of B are estimated



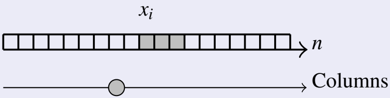
Problem

Finding such zones is hard with:

- continuous speech (non-civilized talk where everyone speaks at the same time)
- short pieces of music

TEMPROM identification stage

1. Successively considering the zones in the above sorted list
2. Estimating a new column (average over the considered zone of the ratio $\frac{x_2}{x_1}$)
3. Keeping it if its distance wrt previously found ones is "sufficiently high"
4. Stoping when all the columns of B are estimated



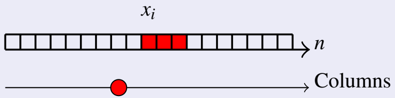
Problem

Finding such zones is hard with:

- continuous speech (non-civilized talk where everyone speaks at the same time)
- short pieces of music

TEMPROM identification stage

1. Successively considering the zones in the above sorted list
2. Estimating a new column (average over the considered zone of the ratio $\frac{x_2}{x_1}$)
3. Keeping it if its distance wrt previously found ones is "sufficiently high"
4. Stoping when all the columns of B are estimated



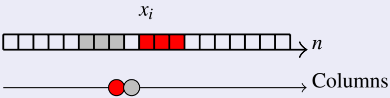
Problem

Finding such zones is hard with:

- continuous speech (non-civilized talk where everyone speaks at the same time)
- short pieces of music

TEMPROM identification stage

1. Successively considering the zones in the above sorted list
2. Estimating a new column (average over the considered zone of the ratio $\frac{x_2}{x_1}$)
3. Keeping it if its distance wrt previously found ones is "sufficiently high"
4. Stoping when all the columns of B are estimated



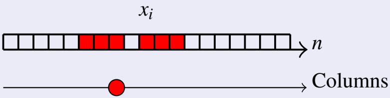
Problem

Finding such zones is hard with:

- continuous speech (non-civilized talk where everyone speaks at the same time)
- short pieces of music

TEMPROM identification stage

1. Successively considering the zones in the above sorted list
2. Estimating a new column (average over the considered zone of the ratio $\frac{x_2}{x_1}$)
3. Keeping it if its distance wrt previously found ones is "sufficiently high"
4. Stoping when all the columns of B are estimated



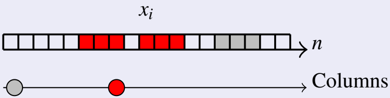
Problem

Finding such zones is hard with:

- continuous speech (non-civilized talk where everyone speaks at the same time)
- short pieces of music

TEMPROM identification stage

1. Successively considering the zones in the above sorted list
2. Estimating a new column (average over the considered zone of the ratio $\frac{x_2}{x_1}$)
3. Keeping it if its distance wrt previously found ones is "sufficiently high"
4. Stoping when all the columns of B are estimated



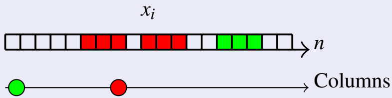
Problem

Finding such zones is hard with:

- continuous speech (non-civilized talk where everyone speaks at the same time)
- short pieces of music

TEMPROM identification stage

1. Successively considering the zones in the above sorted list
2. Estimating a new column (average over the considered zone of the ratio $\frac{x_2}{x_1}$)
3. Keeping it if its distance wrt previously found ones is "sufficiently high"
4. Stoping when all the columns of B are estimated



Problem

Finding such zones is hard with:

- continuous speech (non-civilized talk where everyone speaks at the same time)
- short pieces of music

Increasing sparsity of signals: Frequency analysis

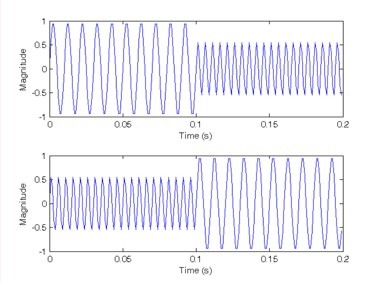
Fourier transform

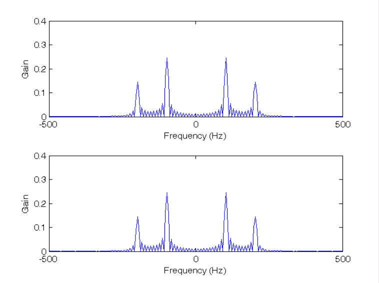
- Joseph Fourier proposed a mathematical tool for computing the frequency information $X(\omega)$ provided by a signal $x(n)$
- Fourier transform is a linear transform:

$$x_1(n) = a_{11}s_1(n) + a_{12}s_2(n) \quad \overset{\text{Fourier transform}}{\overbrace{\longrightarrow}} \quad X_1(\omega) = a_{11}S_1(\omega) + a_{12}S_2(\omega)$$

- Previous TEMPROM approach still applies on Frequency domain.

Limitations of Fourier analysis

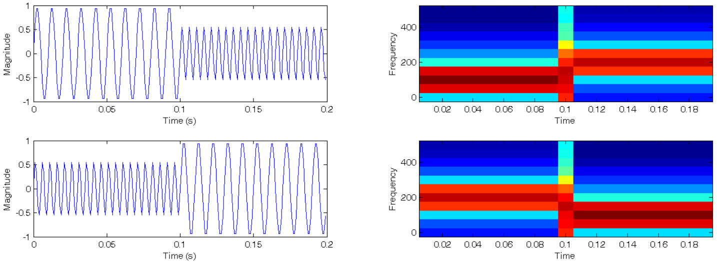




Going further: Time-frequency (TF) analysis

- Musicians are used to TF representations:

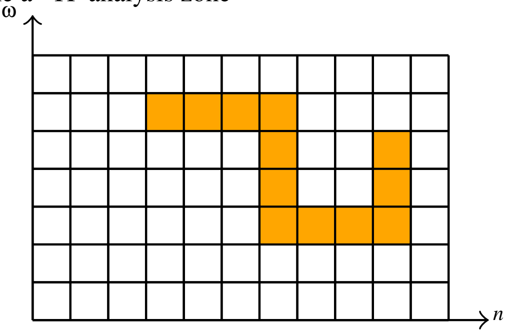
- Short-Term Fourier Transform (STFT):
  1. we cut the signals in small temporal "pieces"
  2. on which we compute the Fourier transform



- $x_1(n) = a_{11}s_1(n) + a_{12}s_2(n) \overset{\text{STFT}}{\longrightarrow} X_1(n,\omega) = a_{11}S_1(n,\omega) + a_{12}S_2(n,\omega)$

From TEMPROM to TIFROM

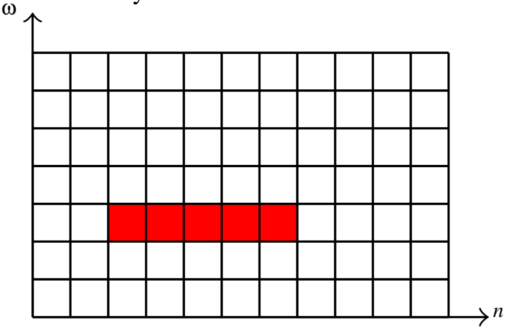
- Extension of the TEMPROM method to **TI**me-**F**requency domain (hence its name TIFROM – Abrard and Deville, 2001–2005)
- Need to define a "TF analysis zone"



- Concept of the approach is then the same
- Further improvements (Deville *et al.*, 2004, and Puigt & Deville, 2009)

From TEMPROM to TIFROM

- Extension of the TEMPROM method to **TI**me-**F**requency domain (hence its name TIFROM – Abrard and Deville, 2001–2005)
- Need to define a “TF analysis zone”



- Concept of the approach is then the same
- Further improvements (Deville *et al.*, 2004, and Puigt & Deville, 2009)

Audio examples

Simulation

- 4 real English-spoken signals
- Mixed with:

$$A = \begin{bmatrix} 1 & 0.9 & 0.9^2 & 0.9^3 \\ 0.9 & 1 & 0.9 & 0.9^2 \\ 0.9^2 & 0.9 & 1 & 0.9 \\ 0.9^3 & 0.9^2 & 0.9 & 1 \end{bmatrix}$$

- Performance measured with signal-to-interference ratio: 49.1 dB

Do you understand something?

Observation 1 Observation 2 Observation 3 Observation 4

Let us separate them:

Output 1 Output 2 Output 3 Output 4

And how were the original sources?

Source 1 Source 2 Source 3 Source 4

Other sparsifying transforms/approximations

Sparsifying approximation

There exists a dictionary Φ such that $s(n)$ is (approximately) decomposed as a linear combination of a few atoms ϕ_k of this dictionary, i.e.
$s(n) = \sum_{k=1}^{K} c(k)\phi_k(n)$ where K is "small"

How to make a dictionary?

1. Fixed basis (wavelets, STFT, (modified) discrete cosine transform (DCT), union of bases (e.g. wavelets + DCT), etc)
2. Adaptive basis, i.e. data-learned "dictionaries" (e.g. K-SVD)

How to select the atoms?

Given $s(n)$ and Φ, find the sparsest vector $\underline{c}$ such that $s(n) \simeq \sum_{k=1}^{K} c(k)\phi_k(n)$

- ℓ^q-based approaches
- Greedy algorithms (Matching Pursuit and its extensions)

Sparsifying transforms useful and massively studied, with many applications (e.g. denoising, inpainting, coding, compressed sensing, etc). Have e.g. a look to Sturm (2009)

Underdetermined case: partial separation / cancellation

Configuration when N sources / P observations

- $P = N$: No problem
- $P > N$: No problem (e.g. dimension reduction thanks to PCA)
- $P < N$: Underdetermined case (more sources than observations) $\Rightarrow$ B non-invertible!

- Estimation of columns of B: same principle than above.
- Partial recovering of the sources

Canceling the contribution of one source

Si S_k **isolated** in an analysis zone: $y_i(n) = x_i(n) - \frac{a_{ik}}{a_{1k}} x_1(n)$.

- *Karaoke*-like application:

  Observation 1 Observation 2 Output "without singer"

- Many audio examples on: `http://www.ast.obs-mip.fr/puigt` (Section: Miscellaneous / secondary school students internship)

Underdetermined case: full separation

- B estimated, perform a full separation $\Rightarrow$ additive assumption
- W-Disjoint Orthogonality (WDO): in each TF window (n, ω), one source is active, which is approximately satisfied for LI speech mixtures (Yilmaz and Rickard, 2004)
  1. Successively considering observations in each TF window (n, ω) and measuring their distance wrt each column of B (e.g. by computing $\frac{X_i(n,\omega)}{X_1(n,\omega)}$)
  2. Associating this TF window with the closest column (i.e. one source)
  3. Creating N binary masks and applying them to the observations
  4. Computing the inverse STFT of resulting signals
- Locally determined mixtures assumption: in each TF window, at most P sources are active (**inverse problems**)

Example (SiSEC 2008):

Observations Source 1 Source 2 Source 3

Binary masking separation: Output 1 Output 2 Output 3

Underdetermined case: full separation

- B estimated, perform a full separation $\Rightarrow$ additive assumption
- W-Disjoint Orthogonality (WDO)
- Locally determined mixtures assumption: in each TF window, at most P sources are active (**inverse problems**)
  1. ℓ^q-norm ($q \in [0,1]$) minimization problems (Bofill & Zibulevsky, 2001, Vincent, 2007, Mohimani *et al.*, 2009)

  $$\min_{\underline{s}} ||\underline{s}||_q \text{ s.t. } \underline{x} = A\underline{s}$$

  2. statistically sparse decomposition (Xiao *et al.*, 2005): In each zone, at most P active sources: $R_{\underline{s}}(\tau) \simeq \begin{bmatrix} R_{\underline{s}}^{sub}{}_{P\times P} & 0_{P\times(N-P)} \\ 0_{(N-P)\times P} & 0_{(N-P)\times(N-P)} \end{bmatrix}$ with $R_{\underline{s}}^{sub} \simeq A_{j_1,\ldots,j_P}^{-1} R_{\underline{x}}(\tau) \left(A_{j_1,\ldots,j_P}^{-1}\right)^T$. Finding them:
  $\left[\widehat{j_1},\ldots,\widehat{j_P}\right] = \arg\min \frac{\sum_{i=1}^{P}\sum_{j>i}\left|R_{\underline{s}}^{sub}(i,j)\right|}{\sqrt{\prod_{i=1}^{P} R_{\underline{s}}^{sub}(i,i)}}$

Example (SiSEC 2008):

Observations　　Source 1　　Source 2　　Source 3

ℓ_p-based separation (Vincent, 2007): Output 1　　Output 2　　Output 3

Conclusion

Conclusion

1. Introduction to a Sparse Component Analysis method
2. Many methods based on the same stages propose improved criteria for finding single-source zones and estimating the mixing parameters
3. General tendency to relax more and more the joint-sparsity assumption
4. Well suited to non-stationary and/or dependent sources
5. Able to process the underdetermined case

LI-TIFROM BSS softwares

`http://www.ast.obs-mip.fr/li-tifrom`

References

- F. Abrard, Y. Deville, and P. White: *From blind source separation to blind source cancellation in the underdetermined case: a new approach based on time-frequency analysis*, Proc. ICA 2001, pp. 734–739, San Diego, California, Dec. 9–13, 2001
- F. Abrard and Y. Deville: *A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources*, Signal Processing, 85(7):1389-1403, July 2005.
- P. Bofill and M. Zibulevsky: *Underdetermined Blind Source Separation using Sparse Representations*, Signal Processing, 81(11):2353–2362, 2001
- Y. Deville, M. Puigt, and B. Albouy: *Time-frequency blind signal separation: extended methods, performance evaluation for speech sources*, Proc. of IEEE IJCNN 2004, pp. 255-260, Budapest, Hungary, 25-29 July 2004.
- H. Mohimani, M. Babaie-Zadeh, and C. Jutten: *A fast approach for overcomplete sparse decomposition based on smoothed L0 norm*, IEEE Transactions on Signal Processing, 57(1):289-301, January 2009.
- M. Puigt and Y. Deville: *Iterative-Shift Cluster-Based Time-Frequency BSS for Fractional-Time-Delay Mixtures*, Proc. of ICA 2009, vol. LNCS 5441, pp. 306-313, Paraty, Brazil, March 15-18, 2009.
- B. Sturm: *Sparse approximation and atomic decomposition: considering atom interactions in evaluating and building signal representations*, Ph.D. dissertation, 2009. `http://www.mat.ucsb.edu/~b.sturm/PhD/Dissertation.pdf`
- E. Vincent: *Complex nonconvex lp norm minimization for underdetermined source separation*,Proc. of ICA 2007, September 2007, London, United Kingdom. pp. 430-437
- M. Xiao, S.L. Xie, and Y.L. Fu: *A statistically sparse decomposition principle for underdetermined blind source separation*, Proc. ISPACS 2005, pp. 165–168, 2005
- O. Yilmaz and S. Rickard: *Blind separation of speech mixtures via time-frequency masking*, *IEEE Transactions on Signal Processing* , 52(7):1830–1847, 2004.

Part II

Independent Component Analysis

This part is partly inspired by F. Theis' online tutorials.
http://www.biologie.uni-regensburg.de/Biophysik/Theis/teaching.html

Probability theory: recallings (1)

- main object: random variable/vector $\underline{x}$
  - definition: a measurable function on a probability space
  - determined by its density $f_{\underline{x}} : \mathbb{R}^P \to [0,1)$
- properties of a probability density function (pdf)
  - $\int_{\mathbb{R}^P} f_{\underline{x}}(\underline{x}) d\underline{x} = 1$
  - transformation: $f_{A\underline{x}}(\underline{x}) = |\det(A)|^{-1} f_{\underline{x}}(A^{-1}\underline{x})$
- indices derived from densities (probabilistic quantities)
  - expectation or mean: $\mathbb{E}(\underline{x}) = \int_{\mathbb{R}^P} \underline{x} f_{\underline{x}}(\underline{x}) d\underline{x}$
  - covariance: $\mathrm{Cov}(\underline{x}) = \mathbb{E}\left\{(\underline{x} - \mathbb{E}\{\underline{x}\})(\underline{x} - \mathbb{E}\{\underline{x}\})^T\right\}$
- decorrelation and independence
  - $\underline{x}$ is decorrelated if $\mathrm{Cov}(\underline{x})$ is diagonal and white if $\mathrm{Cov}(\underline{x}) = I$
  - $\underline{x}$ is independent if its density factorizes $f_{\underline{x}}(x_1, \ldots, x_P) = f_{x_1}(x_1) \ldots f_{x_n}(x_n)$
  - independent $\Rightarrow$ decorrelated (but not vice versa in general)

Probability theory: recallings (2)

- higher-order moments
  - central moment of a random variable $\underline{x} = x$ $(P = 1)$: $\mu_j(x) \triangleq \mathbb{E}\{(x - \mathbb{E}\{x\})^j\}$
  - $\mu_1(x) = \mathbb{E}\{x\}$ mean and $\mu_2(x) = \mathrm{Cov}(x) \triangleq \mathrm{var}(x)$ variance
  - $\mu_3(x)$ is called skewness – measures asymmetry ($\mu_3(x) = 0$ means $\underline{x}$ symmetric)
- kurtosis
  - the combination of moments $\mathrm{kurt}(\underline{x}) \triangleq \mathbb{E}\{\underline{x}^4\} - 3(\mathbb{E}\{\underline{x}^2\})^2$ is called kurtosis of $\underline{x}$
  - $\mathrm{kurt}(\underline{x}) = 0$ if $\underline{x}$ Gaussian, < 0 if sub-Gaussian and > 0 if super-Gaussian (speech is usually modeled by a Laplacian distribution = super-Gaussian)
- sampling
  - in practice density is unknown only some samples i.e. values of random function are given
  - given independent $(x_i)_{i=1,\ldots,P}$ with same density f, then $x_1(\omega), \ldots, x_n(\omega)$ for some event ω are called i.i.d. samples of f
  - strong theorem of large numbers: given a pairwise i.i.d. sequence $(x_i)_{i\in\mathbb{N}}$ in $L^1(\Omega)$, then (for almost all ω)

$$\lim_{P\to+\infty}\left(\frac{1}{P}\sum_{i=1}^{P} x_i(\omega)\right) - \mathbb{E}\{x_1\} = 0$$

Information theory recallings

- entropy
  - $H(\underline{x}) \triangleq -\mathbb{E}_{\underline{x}}\{(\log f_{\underline{x}})\}$ is called the (differential) entropy of $\underline{x}$
  - transformation: $H(A\underline{x}) = H(\underline{x}) + \mathbb{E}_{\underline{x}}\{\log|\det A|\}$
  - given $\underline{x}$ let $\underline{x}_{\text{gauss}}$ be the Gaussian with mean $\mathbb{E}\{\underline{x}\}$ and covariance $\text{Cov}(\underline{x})$; then $H(\underline{x}_{\text{gauss}}) \geq H(\underline{x})$
- negentropy
  - negentropy of $\underline{x}$ is defined by $J(\underline{x}) \triangleq H(\underline{x}_{\text{gauss}}) - H(\underline{x})$
  - transformation: $J(A\underline{x}) = J(\underline{x})$
  - approximation: $J(\underline{x}) \simeq \frac{1}{12}\mathbb{E}\{\underline{x}^3\}^2 + \frac{1}{48}\text{kurt}(\underline{x})^2$
- information
  - $I(\underline{x}) \triangleq \sum_{i=1}^{P}(H(x_i)) - H(\underline{x})$ is called mutual information of X
  - $I(\underline{x}) \geq 0$ and $I(\underline{x}) = 0$ if and only if $\underline{x}$ is independent
  - transformation: $I(\Lambda\Delta\underline{x} + \underline{c}) = I(\underline{x})$ for scaling Δ, permutation Λ, and translation $\underline{c} \in \mathbb{R}^P$

Principal Component Analysis

- principal component analysis (PCA)
  - also called Karhunen-Loève transformation
  - very common multivariate data analysis tools
  - transform data to feature space, where few “main features” (principal components) make up most of the data
  - iteratively project into directions of maximal variance $\Rightarrow$ second-order analysis
  - main application: prewhitening and dimension reduction
- model and algorithm
  - assumption: $\underline{s}$ is decorrelated $\Rightarrow$ without loss of generality white
  - construction:
    - eigenvalue decomposition $\mathrm{Cov}(\underline{x})$: $D = V\mathrm{Cov}(\underline{x})V^T$ with diagonal D and orthogonal V
    - PCA-matrix W is constructed by $W \triangleq D^{-1/2}V$ because

$$
\begin{aligned}
\mathrm{Cov}(W\underline{x}) &= \mathbb{E}\{W\underline{x}\underline{x}^T W^T\} \\
&= W\mathrm{Cov}(\underline{x})W^T \\
&= D^{-1/2}VCov(\underline{x})V^T D^{-1/2} \\
&=
\end{aligned}
$$

  - indeterminacy: unique up to right transformation in orthogonal group (set of orthogonal transformations): If W' is another whitening transformation of X, then $I = \mathrm{Cov}(W'\underline{x}) = \mathrm{Cov}(W'W^{-1}W\underline{x}) = W'W^{-1}W^{-T}W'^T$ so $W'W^{-1} \in O(N)$.

Principal Component Analysis

- principal component analysis (PCA)
  - also called Karhunen-Loève transformation
  - very common multivariate data analysis tools
  - transform data to feature space, where few “main features” (principal components) make up most of the data
  - iteratively project into directions of maximal variance $\Rightarrow$ second-order analysis
  - main application: prewhitening and dimension reduction
- model and algorithm
  - assumption: $\underline{s}$ is decorrelated $\Rightarrow$ without loss of generality white
  - construction:
    - eigenvalue decomposition $\mathrm{Cov}(\underline{x})$: $D = V\mathrm{Cov}(\underline{x})V^T$ with diagonal D and orthogonal V
    - PCA-matrix W is constructed by $W \triangleq D^{-1/2}V$ because

$$\begin{aligned} \mathrm{Cov}(W\underline{x}) &= \mathbb{E}\{W\underline{x}\underline{x}^T W^T\} \\ &= W\mathrm{Cov}(\underline{x})W^T \\ &= D^{-1/2}V\mathrm{Cov}(\underline{x})V^T D^{-1/2} \\ &= D^{-1/2}DD^{-1/2} = I. \end{aligned}$$

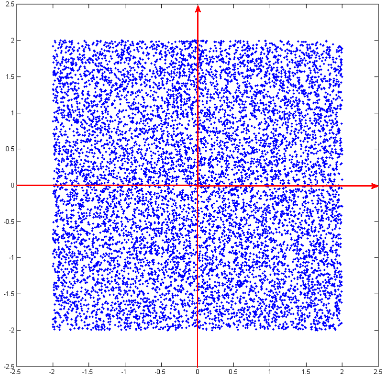
  - indeterminacy: unique up to right transformation in orthogonal group (set of orthogonal transformations): If W' is another whitening transformation of X, then $I = \mathrm{Cov}(W'\underline{x}) = \mathrm{Cov}(W'W^{-1}W\underline{x}) = W'W^{-1}W^{-T}W'^T$ so $W'W^{-1} \in O(N)$.

Algebraic algorithm

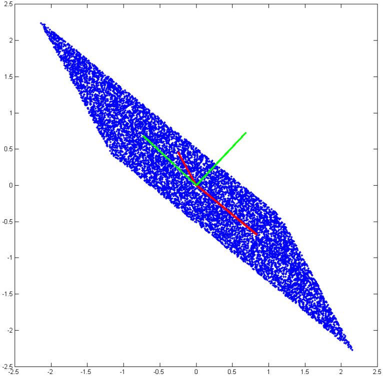
- eigenvalue decomposition
- calculate eigenvectors and eigenvalues of $C \triangleq \mathrm{Cov}(\underline{x})$ i.e. search for $\underline{v} \in \mathbb{R}^P \backslash \{0\}$ with $C\underline{v} = \lambda \underline{v}$
- there exists an orthonormal basis $\{\underline{v}_1, \dots, \underline{v}_P\}$ of eigenvectors of C with corresponding eigenvalues $\lambda_1, \dots, \lambda_P$
- put together we get $V \triangleq [\underline{v}_1 \dots \underline{v}_P]$ and $D \triangleq \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_P \end{bmatrix}$
- hence $CV = VD$ or $V^T CV = D$
- algebraic algorithm
  - in the case of symmetric real matrices (covariance!) construct eigenvalue decomposition by principal axes transformation (diagonalization)
  - PCA-matrix W is given by $W \triangleq D^{-1/2}V$
  - dimension reduction by taking only the N-th ($< P$) largest eigenvalues
- other algorithms (online learning, subspace estimation) exist typically based on neural networks e.g. Oja's rule

From PCA to ICA

- Independent $\Rightarrow$ Uncorrelated (but not the inverse in general)
- Let us see a graphical example with uniform sources$\underline{s}$,$\underline{x} = A\underline{s}$ with $P = N = 2$ and $A = \begin{bmatrix} -0.2485 & 0.8352 \\ 0.4627 & -0.6809 \end{bmatrix}$



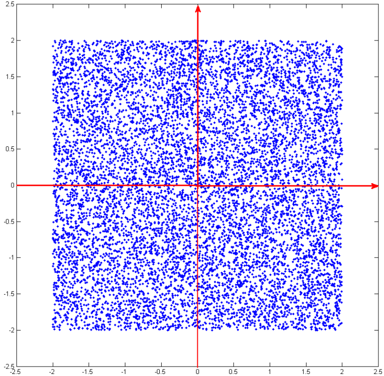
Source distributions (red: source directions)



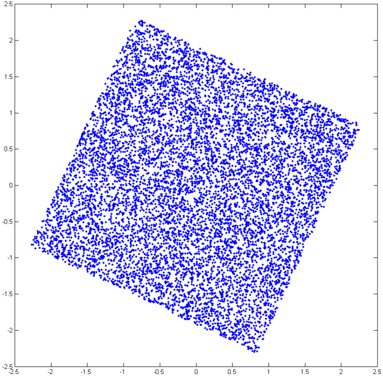
Mixture distributions (green: eigenvectors)

From PCA to ICA

- Independent $\Rightarrow$ Uncorrelated (but not the inverse in general)
- Let us see a graphical example with uniform sources$\underline{s}$,$\underline{x} = A\underline{s}$ with
  $P = N = 2$ and $A = \begin{bmatrix} -0.2485 & 0.8352 \\ 0.4627 & -0.6809 \end{bmatrix}$



Source distributions (red: source directions)



Output distributions after whitening

- PCA does "half the job" and we need to rotate the data to achieve the separation!

- in linear ICA, additional model assumptions are possible
- sources can be assumed to be centered i.e. $\mathbb{E}\{\underline{s}\} = 0$ (coordinate transformation $\underline{x}' \triangleq \underline{x} - \mathbb{E}\{\underline{x}\}$)
- white sources
  - if $A \triangleq [\underline{a}_1 | \ldots | \underline{a}_N]$, then scaling indeterminacy means $\underline{x} = A\underline{s} = \sum_{i=1}^{P} \underline{a}_i s_i = \sum_{i=1}^{P} \left(\frac{\underline{a}_i}{\alpha_i}\right)(\alpha_i s_i)$
  - hence normalization is possible e.g. $\mathrm{var}(s_i) = 1$
- white mixtures (determined case $P = N$):
  - by assumption $\mathrm{Cov}(\underline{s}) = I$
  - let V be PCA matrix of $\underline{x}$
  - then $\underline{z} \triangleq V\underline{x}$ is white, and an ICA of $\underline{z}$ gives ICA of $\underline{x}$
- orthogonal A
  - by assumption $\mathrm{Cov}(\underline{s}) = \mathrm{Cov}(\underline{x}) = I$
  - hence $I = \mathrm{Cov}(\underline{x}) = A\mathrm{Cov}(\underline{s})A^T = AA^T$

ICA algorithms

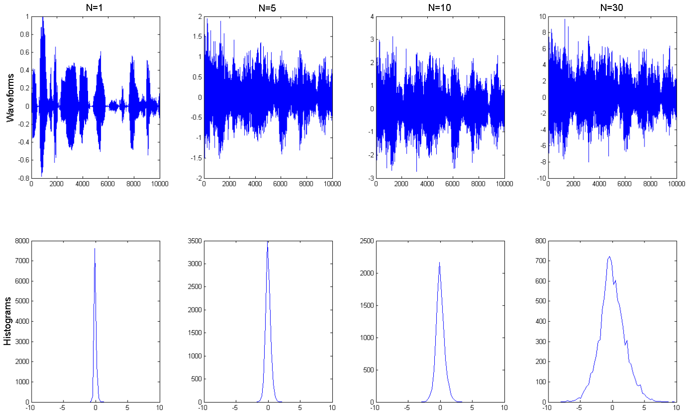
- basic scheme of ICA algorithms (case $P = N$)
- search for invertible $W \in Gl(N)$ that minimizes some dependence measure of WX
  - For example minimize mutual information $I(W\underline{x})$ (Comon, 1994)
  - Or maximize neural network output entropy $H(f(W\underline{x}))$ (Bell and Sejnowski, 1995)
  - Earliest algorithm: extend PCA by performing nonlinear decorrelation (Hérault and Jutten, 1986)
  - Geometric approach, seeing the mixture distributions as a parallelogram whose directions are given by the mixing matrix columns (Theis *et al.*, 2003)
  - Etc...
- We are going to see less briefly:
  - ICA based on non-Gaussianity
  - ICA based on second-order statistics

ICA based on non-Gaussianity

- Mix sources ⇒ Gaussian observations

Why?

Theorem of central limit states that sum of random variables tends to a Gaussian distribution



ICA based on non-Gaussianity

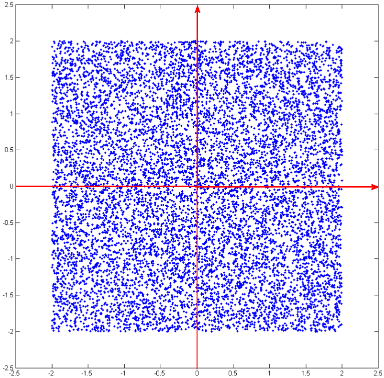
- Mix sources $\Rightarrow$ Gaussian observations

Why?

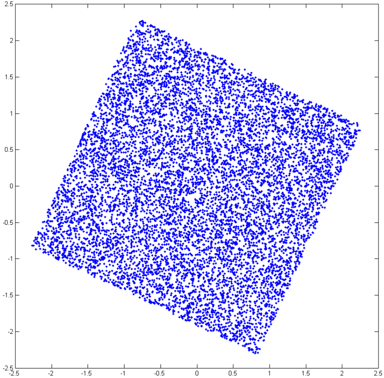
Theorem of central limit states that sum of random variables tends to a Gaussian distribution

- Demixing systems $\Rightarrow$ Non-Gaussian output signals (at most one Gaussian source (Comon, 1994))
- Non-Gaussianity measures:
  - Kurtosis ($\text{kurt}(\underline{x}) = 0$ if $\underline{x}$ Gaussian, > 0 if X Laplacian (speech))
  - Neguentropy (always ≥ 0 and $= 0$ when Gaussian)
- Basic idea: given $\underline{x} = A\underline{s}$, construct ICA matrix W, which ideally equals A^{-1}
  - Recover only one source: search for $\underline{b} \in \mathbb{R}^N$ with $y = \underline{b}^T\underline{x} = \underline{b}^T A\underline{s} \triangleq \underline{q}^T\underline{s}$
  - Ideally $\underline{b}$ is row of A^{-1}, so $\underline{q} = \underline{e_i}$
  - Thanks to central limit theorem $y = \underline{q}^T\underline{s}$ is more Gaussian than all source components s_i
  - At ICA solutions $y \simeq s_i$, hence solutions are *least Gaussian*
- Algorithm (FastICA): Find $\underline{b}$ with $\underline{b}^T\underline{x}$ is maximal non-Gaussian.

Back to our toy example

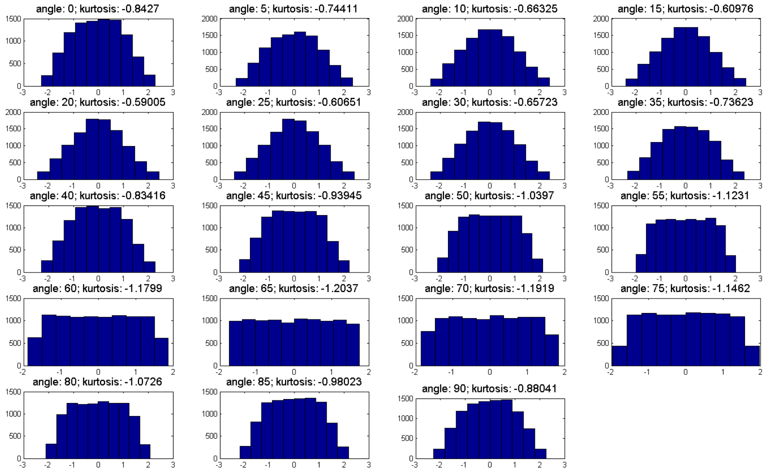


Source distributions



Output distributions after whitening

Back to our toy example

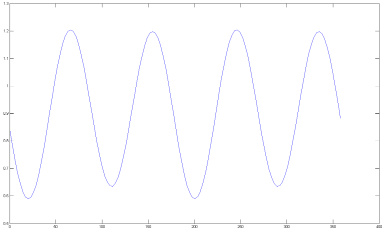


Measuring Gaussianity with kurtosis

- Kurtosis was defined as $\text{kurt}(y) \triangleq \mathbb{E}\{y^4\} - 3\left(\mathbb{E}\{y^2\}\right)^2$
- If y Gaussian, then $\mathbb{E}\{y^4\} = 3\left(\mathbb{E}\{y^2\}\right)^2$, so $\text{kurt}(y) = 0$
- Hence kurtosis (or squared kurtosis) gives a simple measure for the deviation from Gaussianity
- Assumption of unit variance, $\mathbb{E}\{y^2\} = 1$: so $\text{kurt}(y) = \mathbb{E}\{y^4\} - 3$
- two-d example: $\underline{q} = A^T \underline{b} = \begin{bmatrix} q_1 \\ q_2 \end{bmatrix}$
- then $y = \underline{b}^T \underline{x} = \underline{q}^T \underline{s} = q_1 s_1 + q_2 s_2$
- linearity of kurtosis:
  $\text{kurt}(y) = \text{kurt}(q_1 s_1) + \text{kurt}(q_2 s_2) = q_1^4 \text{kurt}(s_1) + q_2^4 \text{kurt}(s_2)$
- normalization: $\mathbb{E}\{s_1^2\} = \mathbb{E}\{s_2^2\} = \mathbb{E}\{y^2\} = 1$, so $q_1^2 + q_2^2 = 1$ i.e. $\underline{q}$ lies on circle

FastICA Algorithm

- $\underline{s}$ is not known $\Rightarrow$ after whitening $underline z = V\underline{x}$ search for $\underline{w} \in \mathbb{R}^N$ with $\underline{w}^T \underline{z}$ maximal non-gaussian
- because of $\underline{q} = (VA)^T \underline{w}$ we get $|\underline{q}|^2 = \underline{q}^T \underline{q} = (\underline{w}^T VA)(A^T V^T \underline{w}) = |\underline{w}|^2$ so if $\underline{q} \in \mathcal{S}^{N-1}$ also $\underline{w} \in \mathcal{S}^{N-1}$
- (kurtosis maximization): Maximize $\underline{w} \mapsto |\mathrm{kurt}(\underline{w}^T \underline{z})|$ on $\mathcal{S}^{N-1}$ after whitening.



$$\phi \mapsto |kurt([cos(\phi), sin(\phi)]\underline{z})|$$

Maximization

- Algorithmic maximization by gradient ascent:
  - A differentiable function $f : \mathbb{R}^N \to \mathbb{R}$ can be maximized by local updates in directions of its gradient
  - Sufficiently small learning rate $\eta > 0$ and a starting point $\underline{x}(0) \in \mathbb{R}^N$, local maxima of f can be found by iterating $\underline{x}(t+1) = \underline{x}(t) + \eta \underline{\Delta x}(t)$ with $\underline{\Delta x}(t) = \operatorname{grad} f(\underline{x}(t)) = \frac{\partial f}{\partial \underline{x}}(\underline{x}(t))$ the gradient of f at $\underline{x}(t)$
- in our case
  $\operatorname{grad}|\operatorname{kurt}(\underline{w}^T \underline{z})|(\underline{w}) = 4\operatorname{sgn}(\operatorname{kurt}(\underline{w}^T \underline{z}))(\mathbb{E}(\underline{z}\{\underline{w}^T \underline{z})^3\} - 3|\underline{w}|^2 \underline{w})$
- Algorithm (gradient ascent kurtosis maximization):
  - Choose $\eta > 0$ and $\underline{w}(0) \in \mathcal{S}^{N-1}$.
  - Then iterate

$$\underline{\Delta w}(t) \triangleq \operatorname{sgn}(\operatorname{kurt}(\underline{w}(t)^T \underline{z}))\mathbb{E}\{\underline{z}(\underline{w}(t)^T \underline{z})^3\}$$

$$\underline{v}(t+1) \triangleq \underline{w}(t) + \eta \underline{\Delta w}(t)$$

$$\underline{w}(t+1) \triangleq \frac{\underline{v}(t+1)}{|\underline{v}(t+1)|}$$

Fixed-point kurtosis maximization

- Local kurtosis maximization algorithm can be improved by this fixed-point algorithm
- any f on $\mathcal{S}^{N-1}$ is extremal at $\underline{w}$ if $\underline{w} \propto \operatorname{grad} f(\underline{w})$
- here: $w \propto \mathbb{E}\{(\underline{w}^T \underline{z})^3 \underline{z}\} - 3|\underline{w}|^2 \underline{w}$
- Algorithm (fixed-point kurtosis maximization): Choose $\underline{w}(0) \in \mathcal{S}^{N-1}$. Then iterate

$$\underline{v}(t+1) \triangleq \mathbb{E}\{(\underline{w}(t)^T \underline{z})^3 \underline{z}\} - 3\underline{w}(t)$$

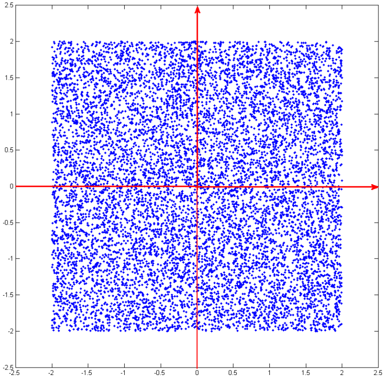
$$\underline{w}(t+1) \triangleq \frac{\underline{v}(t+1)}{|\underline{v}(t+1)|}$$

- advantages:
  - higher convergence speed (cubic instead of quadratic)
  - parameter-free algorithm (apart from the starting vector)
  - $\Rightarrow$ FastICA (Hyvärinen and Oja, 1997)

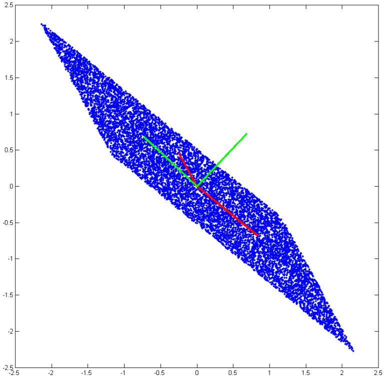
Estimation of more than one component

- By prewhitening, the rows of the whitened demixing W are mutually orthogonal $\rightarrow$ iterative search using one-component algorithm
- Algorithm (deflation FastICA algorithm): Perform fixed-point kurtosis maximization with additional Gram-Schmidt-orthogonalization with respect to previously found ICs after each iteration.
  1. Set $q \triangleq 1$ (current IC).
  2. Choose $\underline{w}_q(0) \in \mathcal{S}^{N-1}$.
  3. Perform a single kurtosis maximization step (here: fixed-point algorithm): $\underline{v}_q(t+1) \triangleq \mathbb{E}\{(\underline{w}_q(t)^T \underline{z})^3 \underline{z}\} - 3\underline{w}_q(t)$
  4. Take only the part of $\underline{v}_p$ that is orthogonal to all previously found $\underline{w}_j$:
  $$\underline{u}_q(t+1) \triangleq \underline{v}_q(t+1) - \sum_{j=1}^{q-1} (\underline{v}_q(t)\, \underline{w}_j)\underline{w}_j$$
  5. Normalize $\underline{w}_q(t+1) \triangleq \frac{\underline{u}_q(t+1)}{|\underline{u}_q(t+1)|}$
  6. If algorithm has not converged go to step 3.
  7. Increment q and continue with step 2 if q is less than the desired number of components.
- alternative: symmetric approach with simultaneous ICA update steps orthogonalization afterwards

Back to our toy example

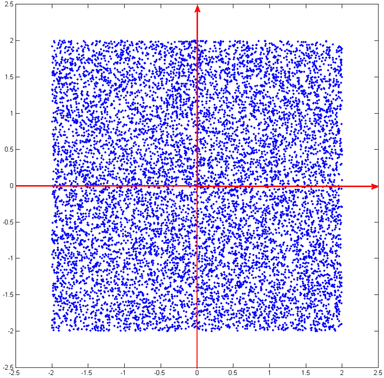


Source distributions (red: source directions)

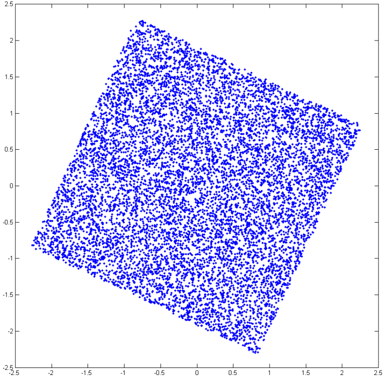


Mixture distributions (green: eigenvectors)

Back to our toy example

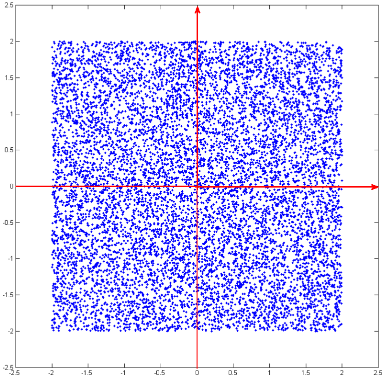


Source distributions (red: source directions)

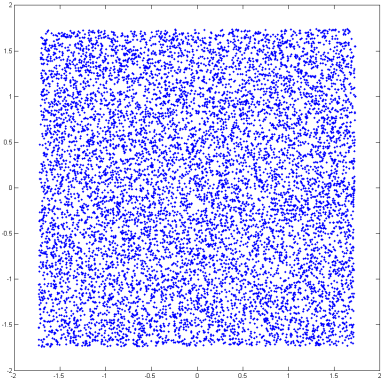


Output distributions after whitening

Back to our toy example



Source distributions (red: source directions)



Output distributions after ICA

ICA based on second-order statistics

- instead of non-Gaussianity of the sources assume here:
  - data possesses additional time structure $\underline{s}(t)$
  - source have diagonal autocovariances $R_{\underline{s}}(\tau) \triangleq \mathbb{E}\{(\underline{s}(t+\tau) - \mathbb{E}\{S(t)\})(S(t) - \mathbb{E}\{S(t)\})\}$ for all τ
- goal: find A (then estimate $\underline{s}(t)$ e.g. using regression)
- as before: centering and prewhitening (by PCA) allow assumptions
  - zero-mean $\underline{x}(t)$ and $\underline{s}(t)$
  - equal source and sensor dimension ($P = N$)
  - orthogonal A
- but hard-prewhitening gives bias...

AMUSE

- bilinearity of autocovariance:

$$R_{\underline{x}}(\tau) = \mathbb{E}\{\underline{x}(t+\tau)\underline{x}(t)^T\} = \begin{cases} AR_{\underline{s}}(0)A^T + \sigma^2 I & \text{if } \tau = 0 \\ AR_{\underline{s}}(\tau)A^T & \text{if } \tau \neq 0 \end{cases}$$

- So symmetrized autocovariance $\overline{R}_{\underline{x}}(\tau) \triangleq \frac{1}{2}\left(R_{\underline{x}}(\tau) + R_{\underline{x}}(\tau)^T\right)$ fulfills (for $\tau \neq 0$)

$$\overline{R}_{\underline{x}}(\tau) = A\overline{R}_{\underline{s}}(\tau)A^T$$

- identifiability:
  - A can only be found up to permutation and scaling (classical BSS indeterminacies)
  - if there exists $\overline{R}_{\underline{s}}(\tau)$ with pairwise different eigenvalues $\Rightarrow$ no more indeterminacies
- AMUSE (algorithm for multiple unknown signals extraction) proposed by Tong *et al.* (1991)
  - recover A by eigenvalue decomposition of $\overline{R}_{\underline{x}}(\tau)$ for one "well-chosen" τ

Other second-order ICA approaches

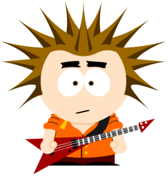
- Limitations of AMUSE:
  - choice of τ
  - susceptible to noise or bad estimates of $\overline{R}_{\underline{x}}(\tau)$
- SOBI (second-order blind identification)
  - Proposed by Belouchrani *et al.* in 1997
  - identify A by joint-diagonalization of a whole set $\{\overline{R}_{\underline{x}}(\tau_1), \overline{R}_{\underline{x}}(\tau_2), \ldots, \overline{R}_{\underline{x}}(\tau_K)\}$ of autocovariance matrices
  - $\Rightarrow$ more robust against noise and choice of τ
- Alternatively, one can assume $\underline{s}$ to be non-stationary
  - hence statistics vary with time
  - joint-diagonalization of non-whitened observations for one (Souloumiac, 1995) or several times τ (Pham and Cardoso, 2001)
- ICA as a generalized eigenvalue decomposition (Parra and Sajda, 2003)
  - 2 lines Matlab code:

```
[W,D] = eig(X*X',R); % compute unmixing matrix W
S = W'*X; % compute sources S
```

  - Discussion about the choice of `R` on Lucas Parra's `quickiebss.html`: `http://bme.ccny.cuny.edu/faculty/lparra/publish/quickiebss.html`

Is audio source independence valid? (Puigt *et al.*, 2009)

Vs.



Naive point of view

- Speech signals are independent...
- While music ones are not!

Signal Processing point of view

It is not so simple...

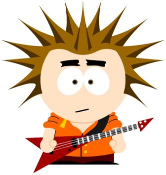
Speech signals
(Smith *et al.*, 2006)

0 — Dependent — Independent → Signal length

Dependent (high correlation) — ???

Music signals
(Abrard & Deville, 2003)

Is audio source independence valid? (Puigt *et al.*, 2009)

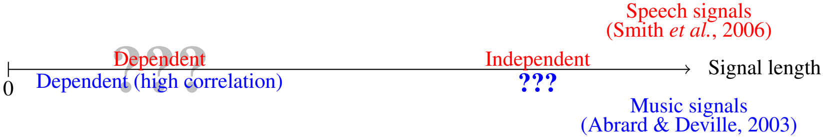


Naive point of view

- Speech signals are independent...
- While music ones are not!

Signal Processing point of view

It is not so simple...



Dependency measures (1)

Tested dependency measures

- Mutual Information (MILCA software, Kraskov *et al.*, 2004):

$$I\{\underline{s}\} = -\mathbb{E}\left\{\log \frac{f_{s_1}(s_1)\ldots f_{s_N}(s_N)}{f_{\underline{s}}(s_1,\ldots,s_N)}\right\}$$

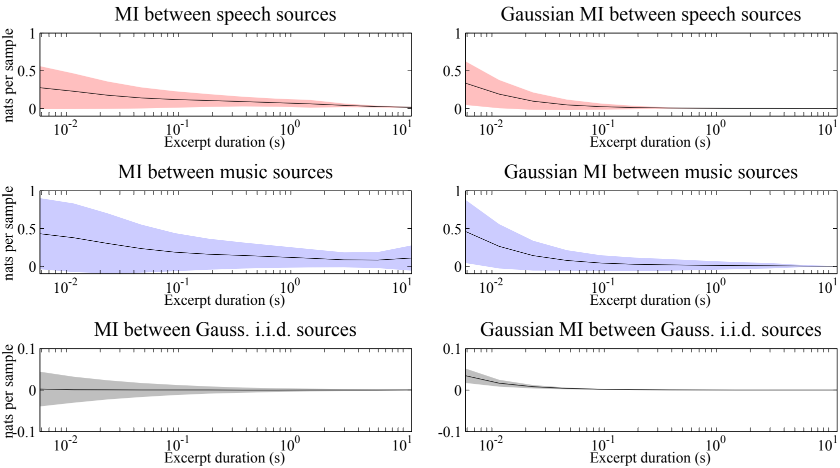
- Gaussian Mutual Information (Pham-Cardoso, 2001):

$$\mathcal{G}I\{\underline{s}\} = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{2}\log\frac{\det\operatorname{diag}\widehat{R}_{\underline{s}}(q)}{\det\widehat{R}_{\underline{s}}(q)}$$

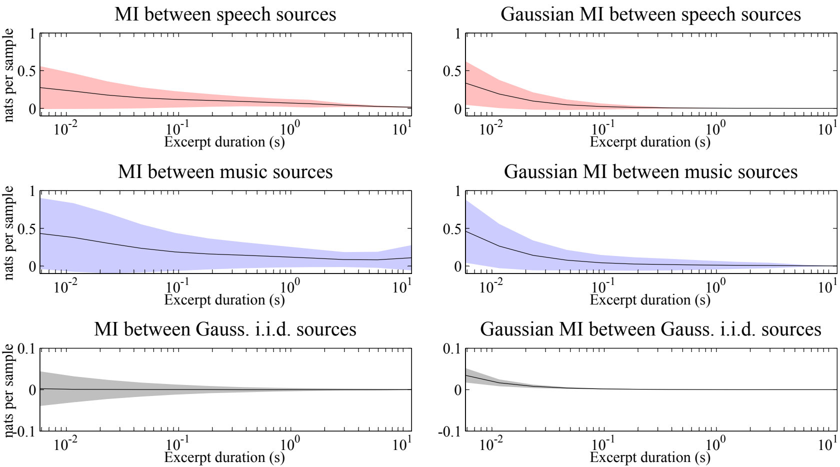
Audio data set

- 90 pairs of signals : 30 pairs of speech + 30 pairs of music + 30 pairs of gaussian i.i.d. signals
- Audio signals cut in time excerpts of 2^7 to 2^{18} samples

Dependency measures (2)



Dependency measures (2)



Speech signals

Dependent
Dependent
(more dependent than speech)

0 → Signal length

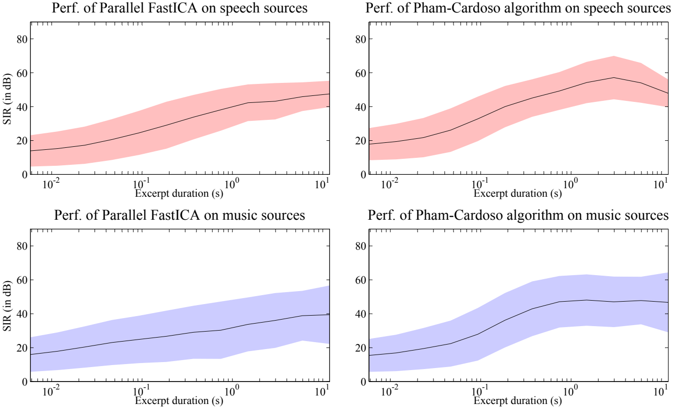
Independent
Independent

Music signals

ICA Performance

Influence of dependency measures on the performance of ICA?

- 60 above pairs (30 speech + 30 music) of audio signals
- Mixed by the Identity matrix
- “Demixed” with Parallel FastICA & the Pham-Cardoso algorithm



ICA Performance

Influence of dependency measures on the performance of ICA?

- 60 above pairs (30 speech + 30 music) of audio signals
- Mixed by the Identity matrix
- "Demixed" with Parallel FastICA & the Pham-Cardoso algorithm

To conclude: behaviour linked to the size of time excerpt

1. High size: same mean behaviour
2. Low size: music signals exhibit more dependencies than speech ones

Conclusion

- Introduction to ICA
- Historical and powerful class of methods for solving BSS
- Independence assumption satisfied in many problems (including audio signals if frames long enough)
- Many criteria have been proposed and some of them are more powerfull than others
- ICA extended to Independent Subspace Analysis (Cardoso, 1998)
- ICA can use extra-information about the sources $\Rightarrow$ Sparse ICA or Non-negative ICA
- Some available softwares:
  - FastICA: `http://research.ics.tkk.fi/ica/fastica/`
  - ICALab: `http://www.bsp.brain.riken.go.jp/ICALAB/`
  - ICA Central algorithms: `http://www.tsi.enst.fr/icacentral/algos.html`
  - many others on personal webpages...

References

- F. Abrard, and Y. Deville: *Blind separation of dependent sources using the "TIme-Frequency Ratio Of Mixtures" approach*, Proc. ISSPA 2003, pp. 81–84.
- A. Bell and T. Sejnowski: *An information-maximisation approach to blind separation and blind deconvolution*, Neural Computation, 7:1129–1159, 1995.
- A. Belouchrani, K. A. Meraim, J.F. Cardoso, and E. Moulines: *A blind source separation technique based on second order statistics*, IEEE Trans. on Signal Processing, 45(2):434–444, 1997.
- J.F. Cardoso: *Blind signal separation: statistical principles*, Proc. of the IEEE, 9(10):2009–2025, Oct. 1998
- P. Comon: *Independent component analysis, a new concept?*, Signal Processing, 36(3):287-314, Apr. 1994
- J. Hérault and C. Jutten: *Space or time adaptive signal processing by neural network models*, Neural Networks for Computing, Proceedings of the AIP Conference, pages 206–211, New York, 1986. American Institute of Physics.
- A. Hyvärinen, J. Karhunen, and E. Oja: *Fast and Robust Fixed-Point Algorithms for Independent Component Analysis*,IEEE Trans. on Neural Networks 10(3)626–634, 1999.
- A. Kraskov, H. Stögbauer, and P. Grassberger: *Estimating mutual information*, Physical Review E, 69(6), preprint 066138, 2004.
- L. Parra and P. Sajda: *Blind Source Separation via generalized eigenvalue decomposition*, Journal of Machine Learning Research, (4):1261–1269, 2003.
- D.T. Pham and J.F. Cardoso: *Blind separation of instantaneous mixtures of nonstationary sources*, IEEE Trans. on Signal Processing, 49(9)1837–1848, 2001.
- M. Puigt, E. Vincent, and Y. Deville: *Validity of the Independence Assumption for the Separation of Instantaneous and Convolutive Mixtures of Speech and Music Sources*, Proc. ICA 2009, vol. LNCS 5441, pp. 613-620, March 2009.
- D. Smith, J. Lukasiak, and I.S. Burnett: *An analysis of the limitations of blind signal separation application with speech*, Signal Processing, 86(2):353–359, 2006.
- A. Souloumiac: *Blind source detection and separation using second-order non-stationarity*, Proc. ICASSP 1995, (3):1912–1915, May 1995.
- F. Theis, A. Jung, C. Puntonet, and E. Lang: *Linear geometric ICA: Fundamentals and algorithms*, Neural Computation, 15:419–439, 2003.
- L. Tong, R.W. Liu, V. Soon, and Y.F. Huang: *Indeterminacy and identifiability of blind identification*, IEEE Transactions on Circuits and Systems, 38:499–509, 1991.

Part III

Non-negative Matrix Factorization

A really really short introduction to NMF

What's that?

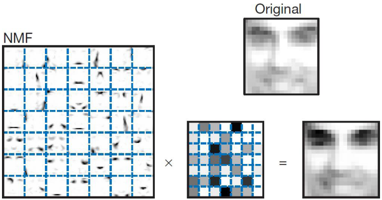
- In many problems, non-negative observations (images, spectra, stocks, etc) $\Rightarrow$ both the sources and the mixing matrix are positive
- Goal of NMF: Decompose a non-negative matrix X as the product of two non-negative matrices A and S with $X = A \cdot S$
- Earliest methods developed in the mid'90 in Finland, under the name of Positive Matrix Factorization (Paatero and Tapper, 1994).
- Famous method by Lee and Seung (1999, 2000) popularized the topic
    - Iterative approach that minimizes the divergence between X and $A \cdot S$

$$\mathrm{div}(X|AS) = \sum_{i,j} \left\{ X_{ij} \log\left[\frac{X_{ij}}{(AS)_{ij}}\right] - X_{ij} + (AS)_{ij} \right\}.$$

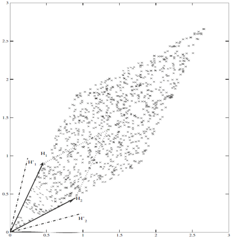
    - Non-unique solution, but uniqueness guaranteed if sparse sources (see e.g. Hoyer, 2004, or Schachtner *et al.*, 2009)
- Plumbley (2002) showed that PCA with non-negativity constraints achieve BSS.

Why is it so popular?

- Face recognition: NMF "recognizes" some natural parts of faces
  - S contains parts-based representation of the data and A is a weight matrix (Lee and Seung, 1999).



- But non-unique solution (see e.g. Schachtner *et al.*, 2009)



Non-negative audio signals? (1)

- If sparsity is assumed, then in each atom, at most one source is active (WDO assumption).
- It then makes sense to apply NMF to audio signals (under the sparsity assumption)

Non-negative audio signals?

- Not in the time domain...
- But OK in the frequency domain, by considering the spectrum of the observations:

$$|\underline{X}(\omega)| = A\,|\underline{S}(\omega)|$$

or

$$|\underline{X}(n,\omega)| = A\,|\underline{S}(n,\omega)|$$

Non-negative audio signals? (2)

- Approaches e.g. proposed by Wang and Plumbley (2005), Virtanen (2007), etc...
- Links between image processing and audio processing when NMF is applied:
  - S now contains a base of waveforms ($\simeq$ a dictionary in sparse models)
  - A still contains a matrix of weights
- Problem: "real" source signals are usually a sum of different waveforms... $\Rightarrow$ Need to do classification after separation (Virtanen, 2007)
- Let us see an example with single observation multiple sources NMF (Audiopianoroll – `http://www.cs.tut.fi/sgn/arg/music/tuomasv/audiopianoroll/`)

Conclusion

- Really really short introduction to NMF
- For audio signals, basically consists in working in the Frequency domain
  - Observations decomposed as a linear combination of basic waveforms
  - Need to cluster them then
- Increasing interest of this family of methods by the community.
- Extensions of NMF to Non-negative Tensor Factorizations.
- More information about the topic on T. Virtanen's tutorial: `http://www.cs.cmu.edu/~bhiksha/courses/mlsp.fall2009/class16/nmf.pdf`
- Softwares:
  - NMFLab: `http://www.bsp.brain.riken.go.jp/ICALAB/nmflab.html`

References

- P. Hoyer: *Non-negative Matrix Factorization with Sparseness Constraints*, Journal of Machine Learning Research 5, pp. 1457–1469, 2004.
- P. Paatero and U. Tapper: *Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values*, Environmetrics 5:111–126, 1994.
- D.D. Lee and H.S. Seung:*Learning the parts of objects by non-negative matrix factorization*, Nature 401 (6755):788–791, 1999.
- D.D. Lee and H.S. Seung:*Algorithms for Non-negative Matrix Factorization.* Advances in Neural Information Processing Systems 13: Proc. of the 2000 Conference. MIT Press. pp. 556–562, 2001.
- M.D. Plumbley: *Conditions for non-negative independent component analysis*, IEEE Signal Processing Letters, 9(6):177–180, 2002
- R. Schachtner, G. Pöppel, A. Tomé, and E. Lang: *Minimum Determinant Constraint for Non-negative Matrix Factorization*, Proc. of ICA 2009, LNCS 5441, pp. 106–113, 2009.
- T. Virtanen: *Monaural Sound Source Separation by Non-Negative Matrix Factorization with Temporal Continuity and Sparseness Criteria*, IEEE Trans. on Audio, Speech, and Language Processing, vol 15, no. 3, March 2007.
- B. Wang and M.D. Plumbley: *Musical audio stream separation by non-negative matrix factorization*, Proc. of the DMRN Summer Conference, Glasgow, 23–24 July 2005.

Part IV

From cocktail party to the origin of galaxies

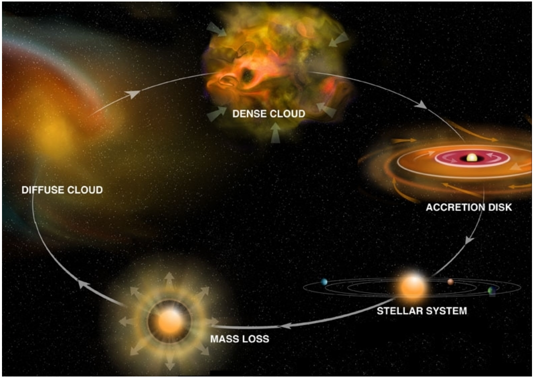
From the paper:
M. Puigt, O. Berné, R. Guidara, Y. Deville, S. Hosseini, C. Joblin: *Cross-validation of blindly separated interstellar dust spectra*, Proc. of ECMS 2009, pp. 41–48, Mondragon, Spain, July 8-10, 2009.

- So far, we focussed on BSS for audio processing.
- But such approaches are generic and may be applied to a much wider class of signals...
- Let us see an example with real data

Problem Statement (1)

Interstellar medium

- Lies between stars in our galaxy
- Concentrated in dust clouds which play a major role in the evolution of galaxies



Adapted from: `http://www.nrao.edu/pr/2006/gbtmolecules/`, Bill Saxton, NRAO/AUI/NSF

Problem Statement (1)

Interstellar medium

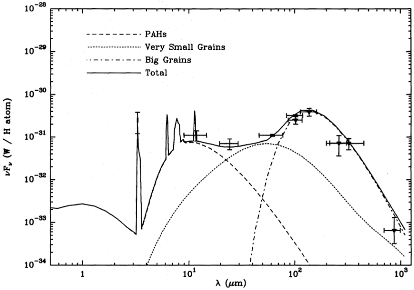
- Lies between stars in our galaxy
- Concentrated in dust clouds which play a major role in the evolution of galaxies

Interstellar dust

- Absorbs UV light and re-emit it in the IR domain
- Several **grains** in Photo-Dissociation Regions (PDRs)
- Spitzer IR spectrograph provides hyperspectral datacubes

$$x_{(n,m)}(\lambda) = \sum_{j=1}^{N} a_{(n,m),j} \, s_j(\lambda)$$

⇨ **Blind Source Separation** (BSS)



- Polycyclic Aromatic Hydrocarbons
- Very Small Grains
- Big grains

Problem Statement (1)

Interstellar medium

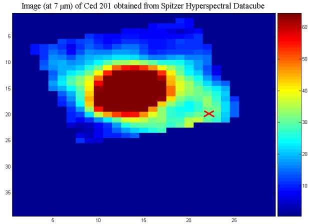
- Lies between stars in our galaxy
- Concentrated in dust clouds which play a major role in the evolution of galaxies

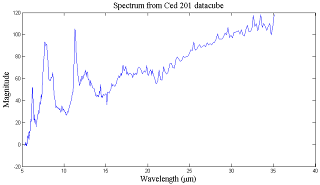
Interstellar dust

- Absorbs UV light and re-emit it in the IR domain
- Several **grains** in Photo-Dissociation Regions (PDRs)
- Spitzer IR spectrograph provides hyperspectral datacubes

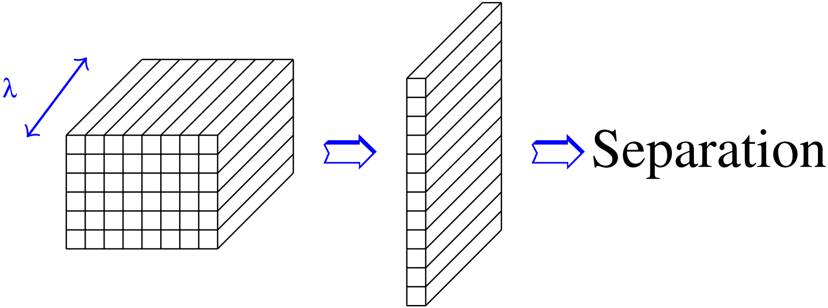
$$x_{(n,m)}(\lambda) = \sum_{j=1}^{N} a_{(n,m),j}\, s_j(\lambda)$$

⇨ **Blind Source Separation** (BSS)





Problem Statement (2)



How to validate the separation of unknown sources?

- Cross-validation of the performance of numerous BSS methods based on different criteria
- Deriving a relevant spatial structure of the emission of grains in PDRs

Blind Source Separation

Three main classes

- **Independent** Component Analysis (ICA)
- **Sparse** Component Analysis (SCA)
- **Non-negative** Matrix Factorization (NMF)

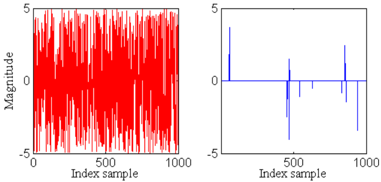
Tested ICA methods

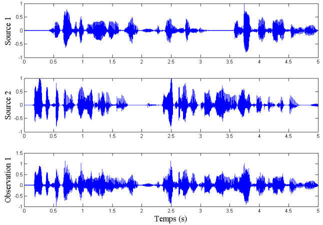
1. FastICA:
   - Maximization of non-Gaussianity
   - Sources are stationary
2. Guidara *et al.* ICA method:
   - Maximum likelihood
   - Sources are Markovian processes & non-stationary

Blind Source Separation

Three main classes

- **Independent** Component Analysis (ICA)
- **Sparse** Component Analysis (SCA)
- **Non-negative** Matrix Factorization (NMF)





Tested SCA methods

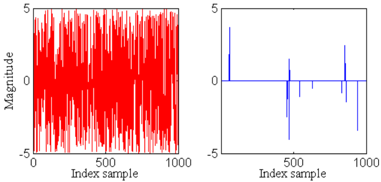
- Low sparsity assumption
- Three methods with the same structure

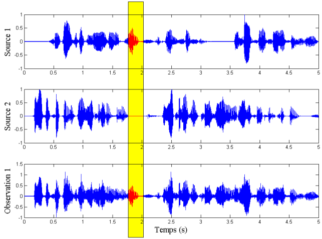
1. LI-TIFROM-S: based on ratios of TF mixtures
2. LI-TIFCORR-C & -NC: based on TF correlation of mixtures

Blind Source Separation

Three main classes

- **Independent** Component Analysis (ICA)
- **Sparse** Component Analysis (SCA)
- **Non-negative** Matrix Factorization (NMF)





Tested SCA methods

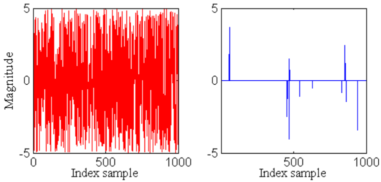
- Low sparsity assumption
- Three methods with the same structure

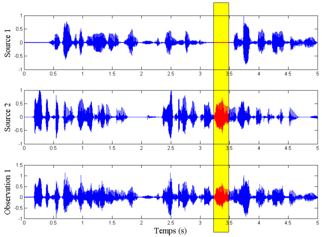
1. LI-TIFROM-S: based on ratios of TF mixtures
2. LI-TIFCORR-C & -NC: based on TF correlation of mixtures

Blind Source Separation

Three main classes

- **Independent** Component Analysis (ICA)
- **Sparse** Component Analysis (SCA)
- **Non-negative** Matrix Factorization (NMF)





Tested SCA methods

- Low sparsity assumption
- Three methods with the same structure

1. LI-TIFROM-S: based on ratios of TF mixtures
2. LI-TIFCORR-C & -NC: based on TF correlation of mixtures

Blind Source Separation

Three main classes

- **Independent** Component Analysis (ICA)
- **Sparse** Component Analysis (SCA)
- **Non-negative** Matrix Factorization (NMF)

Tested NMF method

Lee & Seung algorithm:

- Estimate both mixing matrix $\widehat{A}$ and source matrix $\widehat{S}$ from observation matrix X

Minimization of the divergence between observations and estimated matrices:

$$\operatorname{div}\left(X|\widehat{AS}\right) = \sum_{i,j}\left\{X_{ij}\log\left(\frac{X_{ij}}{\left(\widehat{AS}\right)_{ij}}\right) - X_{ij} + \left(\widehat{AS}\right)_{ij}\right\}$$

Pre-processing stage

- Additive noise not taken into account in the mixing model
- More observations than sources

⇨ Pre-processing stage for reducing the noise & the complexity:

For ICA and SCA methods

1. Sources centered and normalized
2. Principal Component Analysis

For NMF method

- Above pre-processing stage not possible
- Presence of some rare **negative samples** in observations

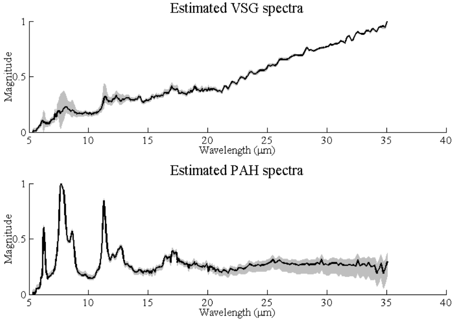
⇨ Two scenarii

1. Negative values are outliers not taken into account
2. Negativeness due to pipeline: translation of the observations to positive values

Estimated spectra from Ced 201 datacube

© R. Croman `www.rc-astro.com`

- Black: Mean values
- Gray: Enveloppe

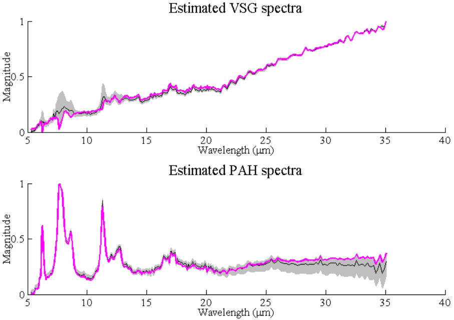


Estimated spectra from Ced 201 datacube

© R. Croman www.rc-astro.com

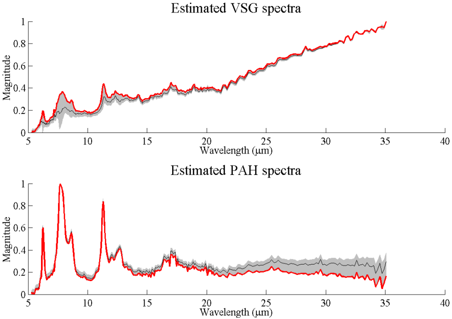
- Black: Mean values
- Gray: Enveloppe

NMF with 1st scenario



Estimated spectra from Ced 201 datacube

© R. Croman www.rc-astro.com



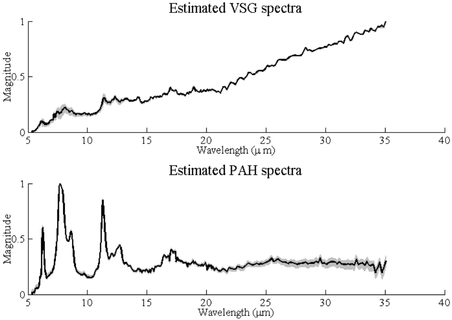
- Black: Mean values
- Gray: Enveloppe

NMF with 1st scenario

FastICA

Estimated spectra from Ced 201 datacube

© R. Croman `www.rc-astro.com`



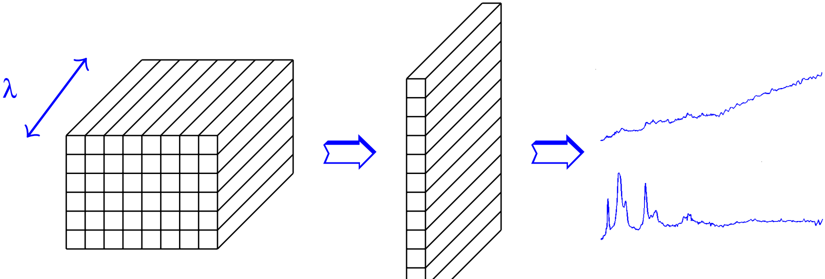
- Black: Mean values
- Gray: Enveloppe

NMF with 1st scenario

FastICA

All other methods

Distribution map of chemical species



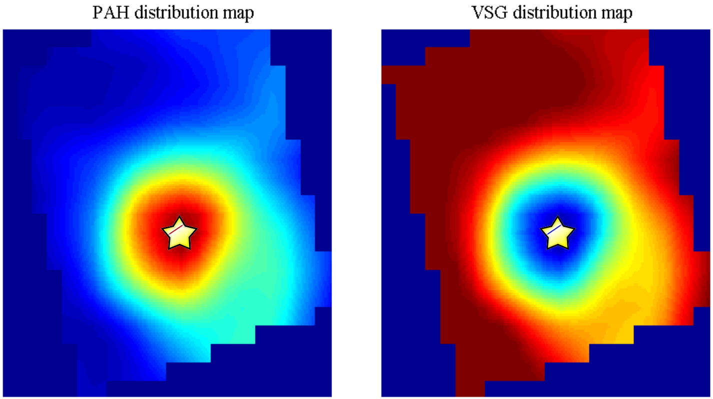
$$x_{(n,m)}(\lambda) = \sum_{j=1}^{N} a_{(n,m),j}\, s_j(\lambda)$$

$$y_k(\lambda) = \eta_j s_j(\lambda)$$

How to compute distribution map of grains?

$$c_{n,m,k} = \mathbb{E}\left\{x_{(n,m)}(\lambda) y_k(\lambda)\right\} = a_{(n,m),j}\, \eta_j \mathbb{E}\left\{s_j(\lambda)^2\right\}$$

Distribution map of chemical species



Conclusion

Conclusion

1. Cross-validation of separated spectra with various BSS methods
   - Quite the same results with all BSS methods
   - Physically relevant
2. Distribution maps provide another validation of the separation step
   - Spatial distribution not used in the separation step
   - Physically relevant