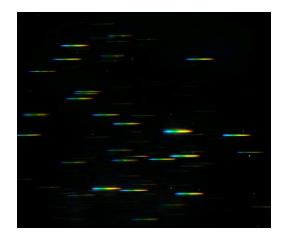






#### Soutenance de stage

## Séparation aveugle des spectres de galaxies



Stagiaire: **Daria MALIK** 26.06.2019 - 30.08.2019

Maître de stage : Shahram HOSSEINI Institut de recherche en astrophysique et planétologie

# Plan

- 1. Présentation de l'IRAP et CNRS
- 2. Contexte du stage
- 3. Sujet de stage Séparation des sources
- 4. Déroulement de stage
- 5. Bilan personnel Conclusion

#### IRAP et CNRS



#### Centre National de la Recherche Scientifique

- établissement public à caractère scientifique et technologique (EPST)
- mène des recherches scientifiques, valorise et partage ses résultats et connaissances
- comprend environ 1 100 laboratoires en France et 36 unités mixtes de recherche internationales



#### Institut de recherche en astrophysique et planétologie

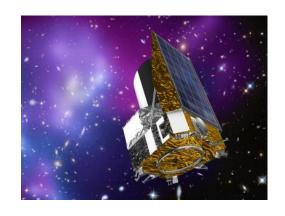
- unité mixte de recherche du CNRS et de l'Université Paul Sabatier
- mène des recherches consacrées à l'étude et la compréhension de l'Univers, développe les projets instrumentaux
- comprend 6 groupes thématiques, environ 300 personnels et étudiants

#### Stage de professionnalisation

Satellite EUCLID (ESA et NASA)

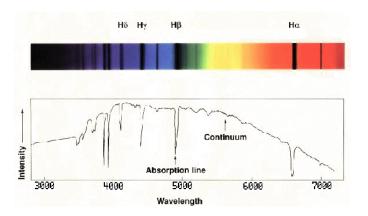


Lancement prévu en Juin 2022 au bord du vaisseau spatial Soyouz

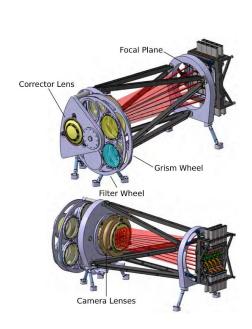




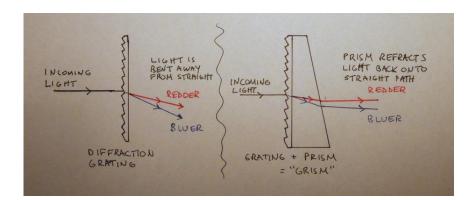
Mission principale: mesurer les spectres de plusieurs millions de galaxies afin de permettre aux scientifiques d'estimer les décalages spectraux (redshift) des galaxies et comprendre plus sur l'expansion de notre Univers et l'énergie noire



#### "Grisme"



Spéctro-photomètre d'EUCLID



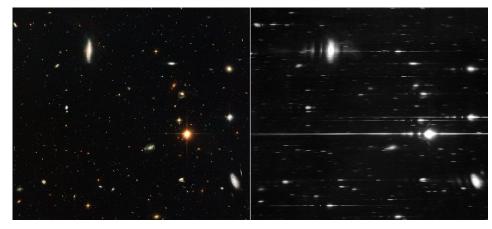
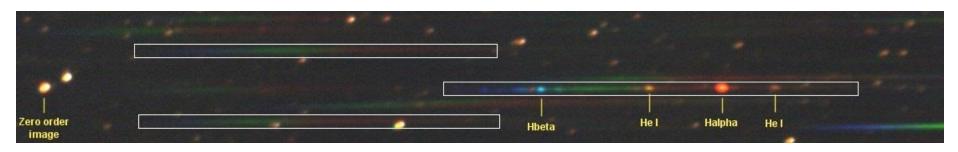


Photo prise par le télescope de Hubble

#### "Grisme"





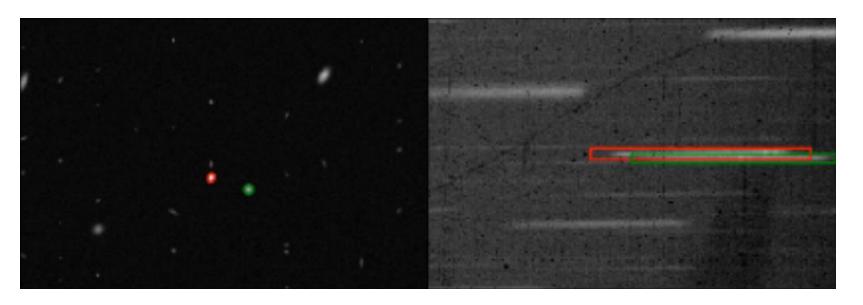
Spectre de la nova Vul avec les raies d'émission d'Hydrogène

## Mélange des spectres

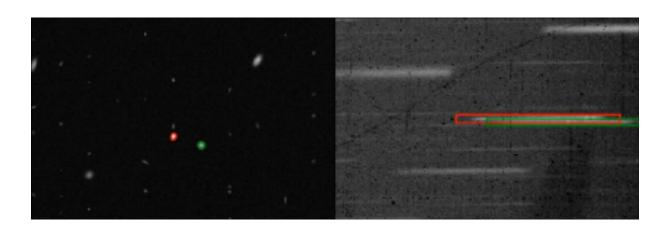


Objets célestes étudiés

Spectres mélangés des objets



#### Stage de professionnalisation





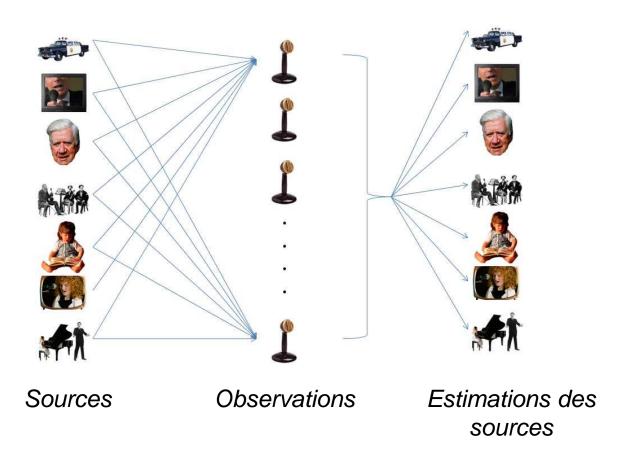
Mettre en oeuvre différents algorithmes de séparation des sources sous MatLab et les appliquer aux données spectrales de galaxies.

Les données sont issues d'un simulateur qui modélise des images représentatives de ce que le télescope du satellite EUCLID renverra au sol lors de sa mission.

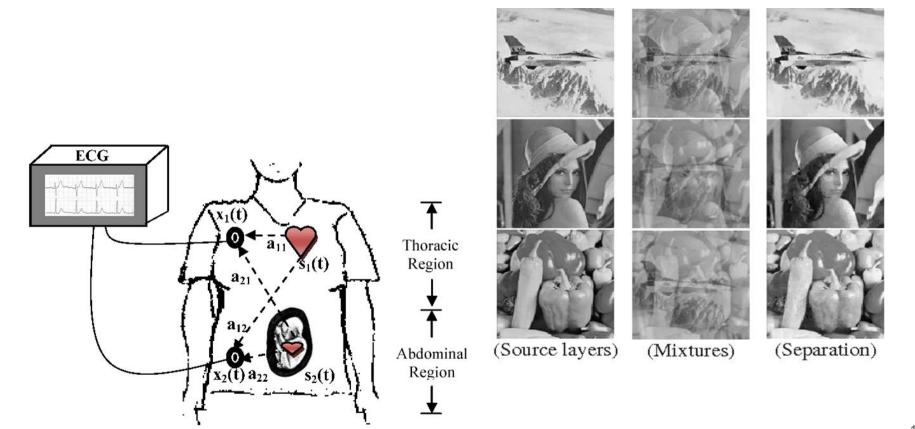
# Plan

- 1. Présentation de l'IRAP et CNRS
- 2. Contexte du stage
- 3. Sujet de stage Séparation des sources
- 4. Déroulement de stage
- 5. Bilan personnel Conclusion

## Séparation des sources ou SAS



## Séparation des sources ou SAS



## Séparation des sources ou SAS

Trois grandes familles des méthodes de SAS :

Analyse en composantes indépendantes (ICA)

*Hypothèse* : *l'indépendance* statistique des sources

<u>Analyse en composantes parcimonieuses (SCA)</u>

*Hypothèse : la parcimonie des sources* 

Décomposition en matrices non-négatives (NMF)

Hypothèse : la non-négativité des mélanges et des sources

# Plan

- 1. Présentation de l'IRAP et CNRS
- 2. Contexte du stage
- 3. Sujet de stage Séparation des sources
- 4. Déroulement de stage
- 5. Bilan personnel Conclusion

## Analyse en composantes indépendantes (ICA)

## Analyse en composantes indépendantes (ICA)

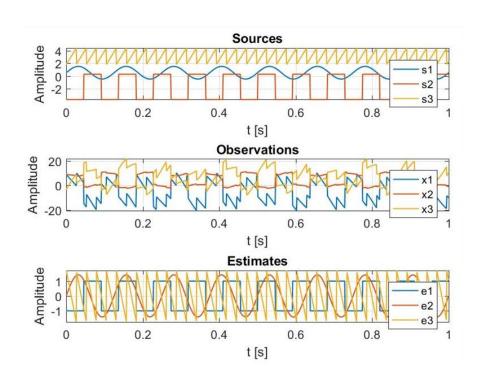
Kurtosis est une mesure de gaussianité. Pour les variables de moyenne nulle il se calcule comme suit  $kurt(y) = E\{y^4\}-3[E\{y^2\}]^2$ . L'algorithme choisi pour optimiser le critère est l'algorithme du gradient.

```
Ewhile 1
54
          y = w1' * X;
55
          y3 = y.^3;
56
57
          X1 = X';
58
          qrad = 4 * (mean([X1(:,1).*y3', X1(:,2).*y3']) - (3*w1'));
59
60
61
          w1 = w1 + (mu*qrad');
62
          w1 = w1 / norm(w1);
63
          if (iter==150)
64
65
              break;
66
          end
67
      end
```

On montre que pour estimer une source le signal y=w.x doit être le moins gaussienne possible.

## Analyse en composantes indépendantes (ICA)

#### Signaux artificiels



#### Signaux audio

9 morceaux de musique mélangés artificiellement en 9 nouveaux morceaux audio

Mix 1 Mix 2

Estimation 1
Estimation 2
Estimation 3

Fonction coût choisi pour mesurer la similitude entre les observations et le produit des matrices estimées A.S est la distance euclidienne  $\frac{1}{2}/|X-AS|/|^2$ 

Fonction coût choisi pour mesurer la similitude entre les observations et le produit des matrices estimées A.S est la distance euclidienne ½//X-AS//²

#### Algorithme du gradient

67

#### *Méthode multiplicative*

#### Alternating Least Squares

```
50
                                                                           S = inv(A'*A)*A'*X;
56
      x = A*S;
                              50
                                     A = A.*((X*S')./(A*S*S'));
57
                              51
                                     A = \max(A, eps);
                                                                           S = max(S, eps);
                                                                   51
      gradA = -(X - X) * S';
58
                              52
                                                                   52
59
      A = A - (mu*gradA);
                              53
                                     S = S.*((A'*X)./(A'*A*S));
                                                                   53
                                                                          A = X*S'*inv(S*S');
60
      A = max(A, eps);
                              54
                                     S = max(S, eps);
                                                                   54
                                                                           A = max(A, eps);
61
                              55
62
      gradS = -A' * (X-x);
                                                                   55
                              56
                                     e = ((X-A*S).^2)./2;
      S = S - (mu*qradS);
63
                                                                   56
                                                                           e = ((X-A*S).^2)./2;
                              57
                                     err = sum(sum(e));
      S = max(S, eps);
64
                                                                   57
                                                                           err = sum(sum(e));
65
66
      e = ((X-A*S).^2)./2;
      err = sum(sum(e));
```

Separation of nonlinear image mixtures nonlinear image mixtures	Separation of nonlinear image mixtures noisuborant. I
Within the area of unsupervised learning, a problem that he been presenting attention is, the note of transforming a set of patterns into new patterns shows compounds, are mountly statistically independent, range Consider, that we are given definencional input data vocators $y=(x_1,y_2)=c$ beeying as probability distribution with density $p_i$ . In general, the dyactions, compounds $x_i$ -only the class within the satisfically interdependent. The problem that we wish to address consists of finding output redevention at one property of the problem that we wish to address consists of finding output redevention at the problem.	Within the area of unsupervised learning and no boaring again and anomalous baning a loagarin as gritiups and W transforming a set of patterns into no "Bering State and the Bering State and the Berich State and the Bering State and the Bering State and the Beri
such that the output components $y_i$ are initially independently in word, a segain langing on the other parameters of the components $y_i$ are initially independently in word, a segain langing on the other parameters of the components $y_i$ and $y_i$ when $y_i$ is inventible, we are simply recoding the data, without any loss of informations if $M_i \leq M$ we are reducing the amount of information present in the data, in the latter case, we wantly wish to ensure that the extracted features $y_i$ are the most important ones, in some appropriate sense, rather at that that $1$ and $1$ are instanced features.	corresponding pairs of points from the two pages of a printed document. The scatter plot of the original images, shown in the battom forward like $a_i$ is the scatter plot of the original images, shown in the battom forward like $a_i$ is the scatter plot of the original images, shown in the battom forward plot $a_i$ is the scatter plot of the plot interest plot of the plot o
right corner, twelvic organized and his hologinal with the process of the process leading from the sources to the	right corner (which corresponds to the lighter intensities in both images) will add search this way reque sidt at indicates that, for those intensities, the mixture is close to singular. Finally, the fact that, the discrete levels of Fig. 2 became largely, Nigreching Fig. 1 is the due to noise in the process. The process leading from the sources to the
i.e. the probability density can be factored into a product of the imaginal dentities of the conquire configurations of the state of the conquire configuration of the state o	observations any observations are the masses on hot baiders of a spect of concern and on a 1 to most office and 1 to the state of the special
There are several reasons for the growing interest that independent component analysis has been beginning in the country in the growing in the property years:  - Recent years:	uniformity in the onion skin paper, especially in its transparency.
The man three states the second of the secon	The murbolle of september 18 feb. of 1900 member 1800 marked 1800 member 1800
himsepments, we that, we obtain will sepincide with the original, a. We will discuss this issue abrades with all comments of the commentation this is the fact, mentiored, the commentation than the original components were able to a commentation than the original components were able to the commentation than the original components were able to the commentation than the original components were able to the commentation than the original components with a commentation that the comm	In this example we intriverential littraties inderinvolve inditartal images, lipituited feet, and graphs "Thin Secial characteristics for interests and graphs in that they not many a work of interests of evidential substitute and an although, due to the above mentioned noise, these, will appears in the personal magical properties of interestry levels, not turnessome tracibilities arom a range of and to be provided and or a call of the substitute of the personal properties and the substitute of the personal properties of the pers
<ul> <li>If the data are to be stored or transmitted, the y are a more efficient representation," because the lon rectandancy ductor the satisticical functorpendence of the components has been eliminated to notify the property of the</li></ul>	In the data are to be stored or transmitted, the year or a more meaning the separation of pipeline properties are always a personal properties of the separation of pipeline properties and the separation of pipeline properties and the separation of properties and the separation of properties and the separation of properties and the properties and the properties and the properties are properties and the properties and the properties are properties are properties and the properties are properties are properties and the properties are properties are properties are properties and the properties are properties are properties and the properties are properties are properties are properties and the properties are properties are properties are properties and the properties are properties and the properties are properties are properties and the propered are properties are properties are properties are properties ar
In the following subsections we shall discuss in some more detail, the linear and nonlinear cases, respectively at all gunzal many and populated and the following subsections we shall discuss in some more detail, the linear and nonlinear cases, respectively at all gunzal many forms to such more than the subsections of the property o	sources can strongly simplify the separation process.  In the case of grayscan mixtures, the use of all challed the first of all challed the gray single of model of the physical mixing.
process should yield much better results that the use of a generic nonlinear separation method. A physical model could have a small number of parameters to be estimated, and would thus allow a much more precise estimation. Furthermore, it might avoid the inherent ill-posedness of nonlinear blind separation. Which is	process should yield much better results that the use of a generic nonlinear separation method. A physical model could have a small number of parameters to be estimated, and would thus allow a much more precise estimation. Furthermore, it might avoid the inherent ill-posedness of nonlinear blind separation, which is
re, ved butting to bloom, a class for a reasonable ved and contact the departure of the case of the function of the case of th	currently addressed through regularization. The parameters of such a model could be estimated by an independent confident advantages are (1) in 1) to the (3) in a south and in 3) and in the parameter of such as a continual and in a confident
documents than statistical independence. In fact, images and/or text from the opposite pages of a printed document can easily happen not to be independent from the conference of the conference	documents than statistical independence. In fact, images and/or text from the opposite pages of a printed document can easily happen not to be independent from one of the opposite pages of landscapes tend to
be igner on the top than on the bottom, inducing a correlation between intensities of both Als as \$1500 of 5500 with with regularly spaced lines, the lines from both sides of the paper may happen for fail on top of each other; or thousands with regularly spaced lines, the lines from both sides of the paper may happen for the one of the other of the fail of the paper may happen for the other of the other oth	be lighter on the top than on the bottom, inducing a correlation between intensities of both. Also, in printed less, with regularly spaced lines, the lines from both sides of the paper may happen to fall on top of each other, or the lines from both sides of the paper may happen to fall on top of each other, or the lines from the lines
could lead to the recovery of the original isources. An important reference in this domain is a paper by Comon	congelation between intensities from both sides of the document. It would be interesting to use or reital based one

In that paper he proved that, in the linear case, the loriginal sources can be recovered if at most one of them has

a Gaussian distribution (all other sources being non-Gaussian). The sourceso zycem appear in y permaned and

a notion of image complexity, but these may not be easy to define, and may be given hardle to use as criterial of in

a Gaussian distribution (all other sources being non-Gaussian). The sourcestyce moistribution (all other sources being non-Gaussian).

#### 1. Introduction

Within the area of unsupervised learning, a problem that has been receiving increasing attention is the one of transforming a set of patterns into new patterns whose components are mutually statistically independent.

Consider that we are given d-dimensional input data vectors  $\mathbf{x} = (x_1, x_2, \dots, x_d)$  obeying a probability distribution with density  $p_x$ . In general, the various components  $x_i$  of the data will be statistically interdependent. The problem that we wish to address consists of finding output vectors

$$y = (y_1, y_2, \dots, y_{d'}) = f(x)$$
 (1)

such that the output components y, are mutually independent.

If d' = d and f is invertible, we are simply recoding the data without any loss of information. If d' < d we are reducing the amount of information present in the data. In the latter case, we usually wish to ensure that the extracted features  $y_i$  are the most important ones, in some appropriate sense.

In this paper we will discuss the first situation, d' = d. If the output components are independent, then

$$p_{\mathbf{y}}(\mathbf{y}) = \prod_{i=1}^{n} p_{y_i}(y_i)$$
 (2)

i.e., the probability density can be factored into a product of the marginal densities of the output components. If we assume that the data x result from a linear combination of independent components, then we can restrict

There are several reasons for the growing interest that independent component analysis has been receiving in

. It can afford a means to perform source separation. Assuming that the observed data x result from an unknown transformation of independent variables z, i.e.

$$g(\mathbf{z})$$

where the  $z = (z_1, z_2, \dots, z_d)$  are unknown source, one may ask whether the independent output components y, that we obtain will coincide with the original z. We will discuss this issue ahead,

- · Related to this is the fact, verified in practice, that the output components y often have a simpler, more intuitive interpretation than the original components  $x_i$ .
- If the data are to be stored or transmitted, the y, are a more efficient representation, because the
  - redundancy due to the statistical interdependence of the components has been eliminated.
- · Humans and animals, as well as artificial systems operating in complex environments, often have to estimate probabilities of events. If a factorial representation is found, then according to Eq. 2 only the marginal densities p., are needed to estimate these probabilities. Otherwise, the much more complex joint density has to be stored.
  - It has been argued that the brains of humans and animals often perform factorial coding operations.
- In the following subsections we shall discuss in some more detail the linear and nonlinear cases, respectively.

#### 1.1. Linear Case

In this case the functions g in (3) and f in (1) are assumed to be linear. The components of the observed data,  $x_i$ , are linear mixtures (with unknown weights) of the original sources  $z_i$ . If we consider the vectors as column matrices, we can write

$$\mathbf{x} = \mathbf{G}\mathbf{z}$$
  
 $\mathbf{y} = \mathbf{F}\mathbf{x}$ 

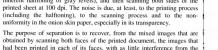
where G and F are  $d \times d$  matrices.

Jutten and Hérault2 were the first to show that this problem could be solved. They also showed that the solution could lead to the recovery of the original sources. An important reference in this domain is a paper by Comon<sup>3</sup>. In that paper he proved that, in the linear case, the original sources can be recovered, if at most one of them has a Gaussian distribution (all other sources being non-Gaussian). The sources z, can appear in y permuted and

#### Separation of nonlinear image mixtures

When acquiring an image of a printed document, the image printed on the opposite page often shows through, due to partial transparency of the paper. Here we are dealing with quite a strong case of that effect, because we're using onion skin paper which is quite transparent.

The mixture that is obtained is rather nonlinear, as can be observed from the top figure on the right, which shows a scatter plot of the intensities of corresponding pairs of points from the two pages of a printed document. The scatter plot of the original images, shown in the bottom figure, filled a square, and had only a relatively small number of discrete intensity levels for each image. The fact that the shape of the scatter plot of Fig. 1 is very different from a parallelogram shows that the mixture was strongly nonlinear. The fact that this scatter plot becomes quite narrow in the upperright corner (which corresponds to the lighter intensities in both images) indicates that, for those intensities, the mixture is close to singular. Finally, the fact that the discrete levels of Fig. 2 became largely blurred in Fig. 1 is due to noise in the process. The process leading from the sources to the observations involved printing the images, on both sides of a sheet of onion skin paper, at 1200 dpi, with a black and white laser printer (with the inherent halftoning of gray levels), and then scanning both sides of the



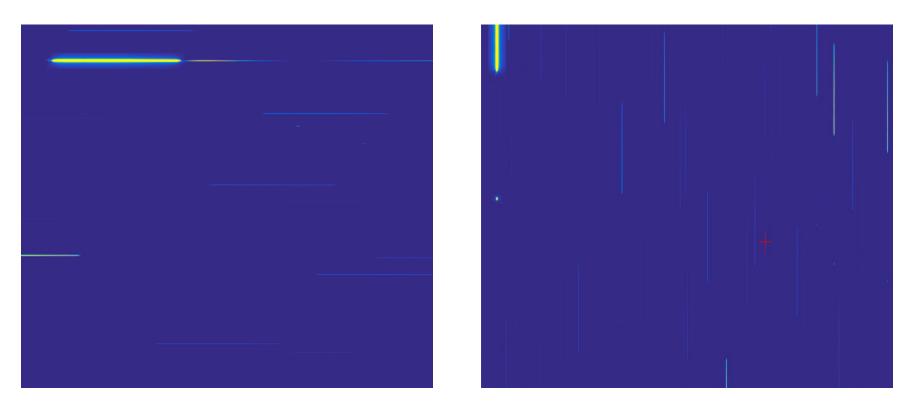
The purpose of separation is to recover, from the mixed images that are obtained by scanning both faces of the printed document, the images that had been printed in each of its faces, with as little interference from the other image as possible.

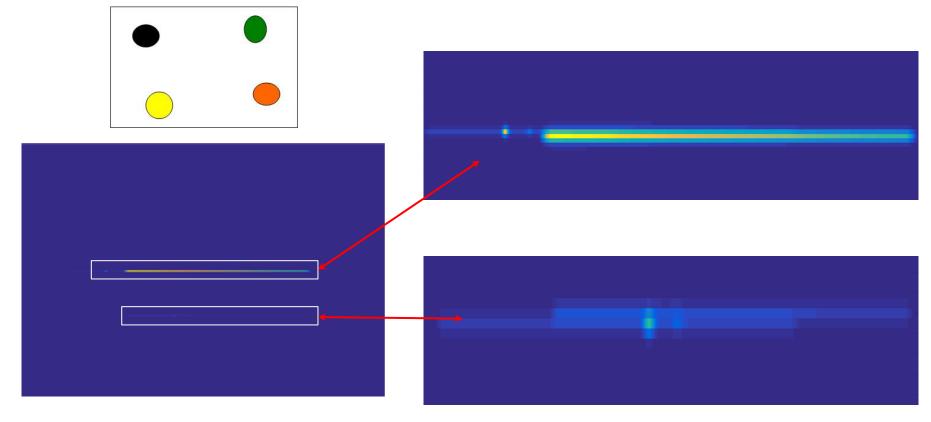
In this example we are creating mixtures that involve natural images, printed text and graphs. The special characteristic of printed text and graphs is that they normally involve just two intensity levels (black and white) although, due to the above mentioned noise, these will appear, in the scanned images, as two clusters of intensity levels.

The separation of mixtures of two-level images, such as printed text, may be much easier than the separation of grayscale images. In fact, at least in the case of mixtures that are not too strong, a simple thresholding procedure may yield the desired results. Such a procedure can be easily performed by hand with most image processing programs, and should not be hard to automate. In such a case the use of more general blind source separation methods might be an overkill, both because it would involve a much larger amount of processing and because it might actually yield worse results. This is an extreme case in which prior knowledge about the sources can strongly simplify the separation process.

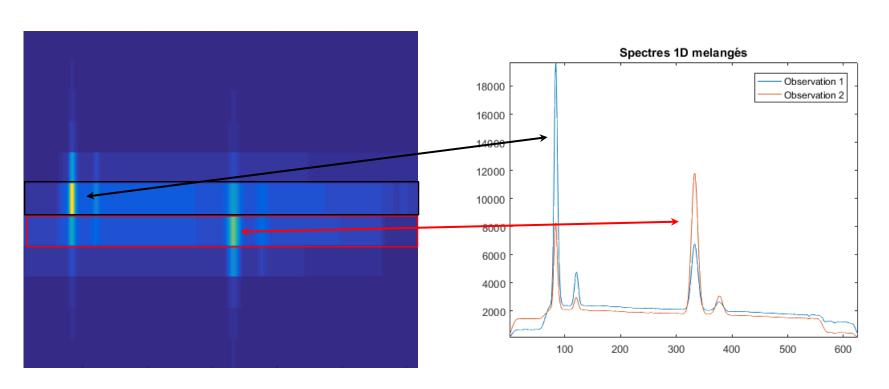
In the case of grayscale mixtures, the use of a separation method based on a good model of the physical mixing process should yield much better results that the use of a generic nonlinear separation method. A physical model could have a small number of parameters to be estimated, and would thus allow a much more precise estimation. Furthermore, it might avoid the inherent ill-posedness of nonlinear blind separation, which is currently addressed through regularization. The parameters of such a model could be estimated by an independent component analysis criterion.

Another issue of interest is the definition of separation criteria that are more suited for images or for printed documents than statistical independence. In fact, images and/or text from the opposite pages of a printed document can easily happen not to be independent from one other. For examples, images of landscapes tend to be lighter on the top than on the bottom, inducing a correlation between intensities of both. Also, in printed text with regularly spaced lines, the lines from both sides of the paper may happen to fall on top of each other, or the lines from one side may fall on the intervals of the lines from the other side, also inducing a significant correlation between intensities from both sides of the document. It would be interesting to use criteria based on a notion of image complexity, but these may not be easy to define, and may be even harder to use as criteria for optimizing a source separation system.

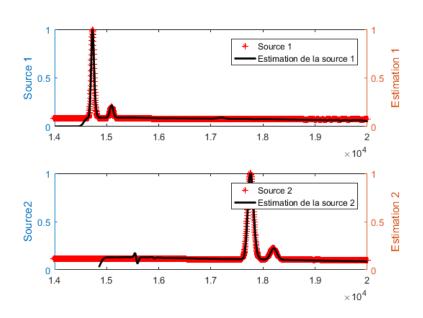




1 image - 2 sources mélangées



1 image - 2 sources mélangées



Source 1 Source 1 Estimation de la source Estimation 1.5 1.6 1.7 1.8 1.9  $\times 10^4$ Source 2 Estimation 2 Estimation de la source 2 Source<sub>2</sub> 1.5 1.6 1.7 1.8 1.9 1.4 2  $\times 10^4$ 

NMF, méthode multiplicative

ICA, algorithme du gradient

# Plan

- 1. Présentation de l'IRAP et CNRS
- 2. Contexte du stage
- 3. Sujet de stage Séparation des sources
- 4. Déroulement de stage
- 5. Bilan personnel Conclusion

#### Bilan des compétences et connaissances





#### approfondies et mises en oeuvre

codage sous MatLab

notions de base de traitement du signal et des images

organisation du temps et des priorités des tâches

communication et intégration

recherche d'informations dans les sources académiques

#### acquises

méthodes de Séparation des sources

notions théoriques en spectroscopie spatiale

algorithmes d'optimisation

#### Bilan personnel



Travail plus individuel mais interactions fréquentes avec les autres stagiaires et membres de l'équipe SISU ⇒ échanges constructifs et bonne ambiance



Difficulté d'organiser le temps suivant les priorités des tâches et les deadlines ⇒ recherche de nouvelles techniques de timemanagement



Découvert d'un nouveau champ de connaissances et gain d'expériences ⇒ évolution personnelle et professionnelle