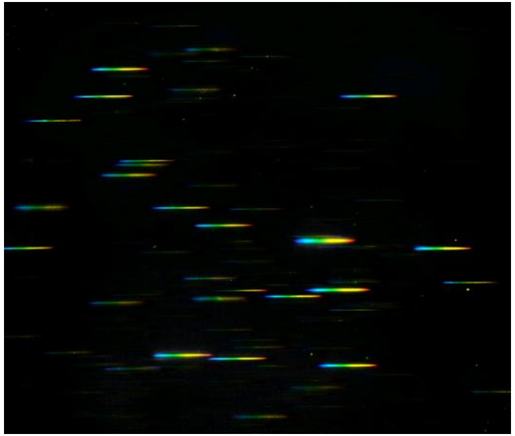
Soutenance de stage

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

Licence 3 EEA CMI - Université Toulouse III Paul Sabatier

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

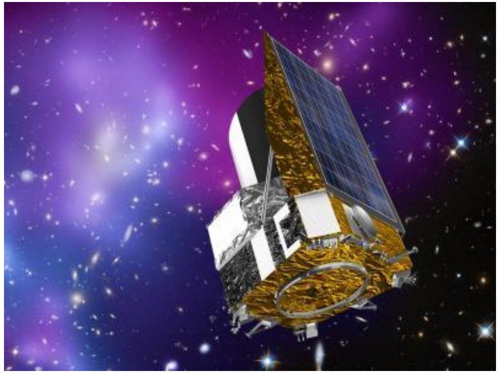
- établissement public à caractère scientifique et technologique (EPST)
- mène des recherches scientifiques, valorise et partage ses résultats et connaissances
- comprend environ 1 100 laboratoires en France et 36 unités mixtes de recherche internationales

Institut de recherche en astrophysique et planétologie

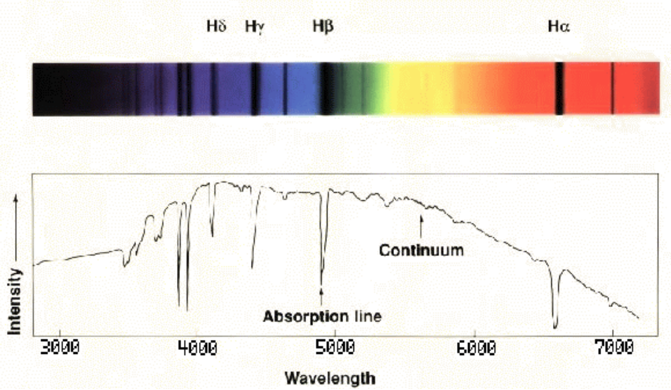
- unité mixte de recherche du CNRS et de l'Université Paul Sabatier
- mène des recherches consacrées à l'étude et la compréhension de l'Univers, développe les projets instrumentaux
- comprend 6 groupes thématiques, environ 300 personnels et étudiants

Stage de professionnalisation

Satellite EUCLID (ESA et NASA)

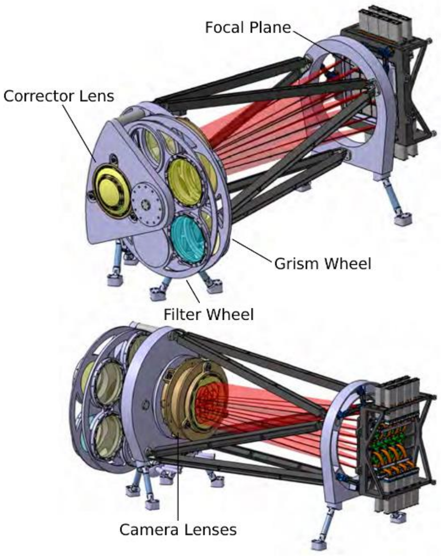


Lancement prévu en Juin 2022 au bord du vaisseau spatial Soyouz



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

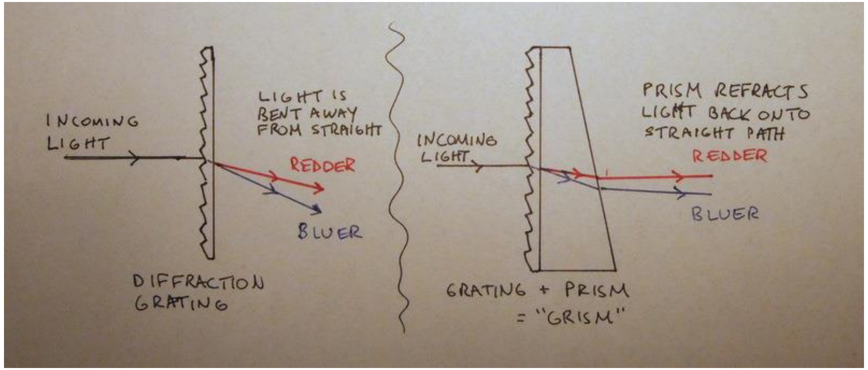
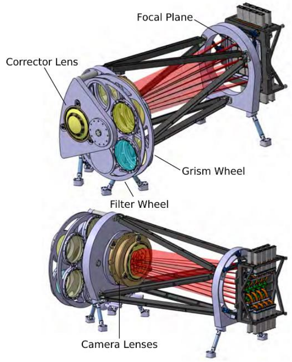
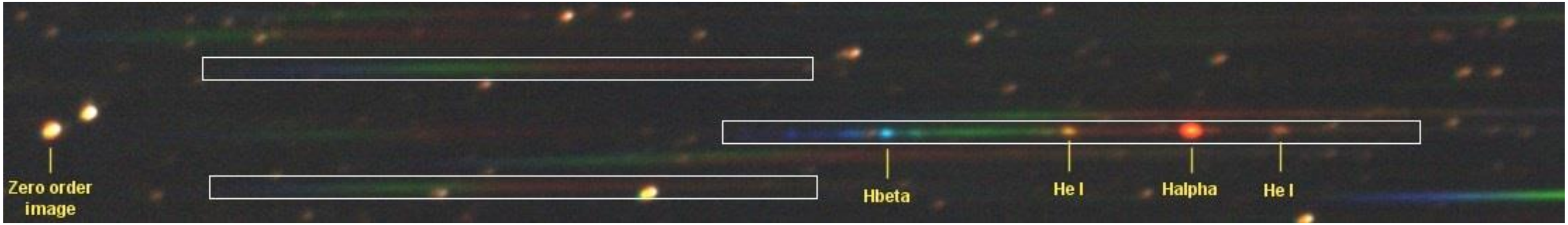


Photo prise par le télescope de Hubble

“Grisme”



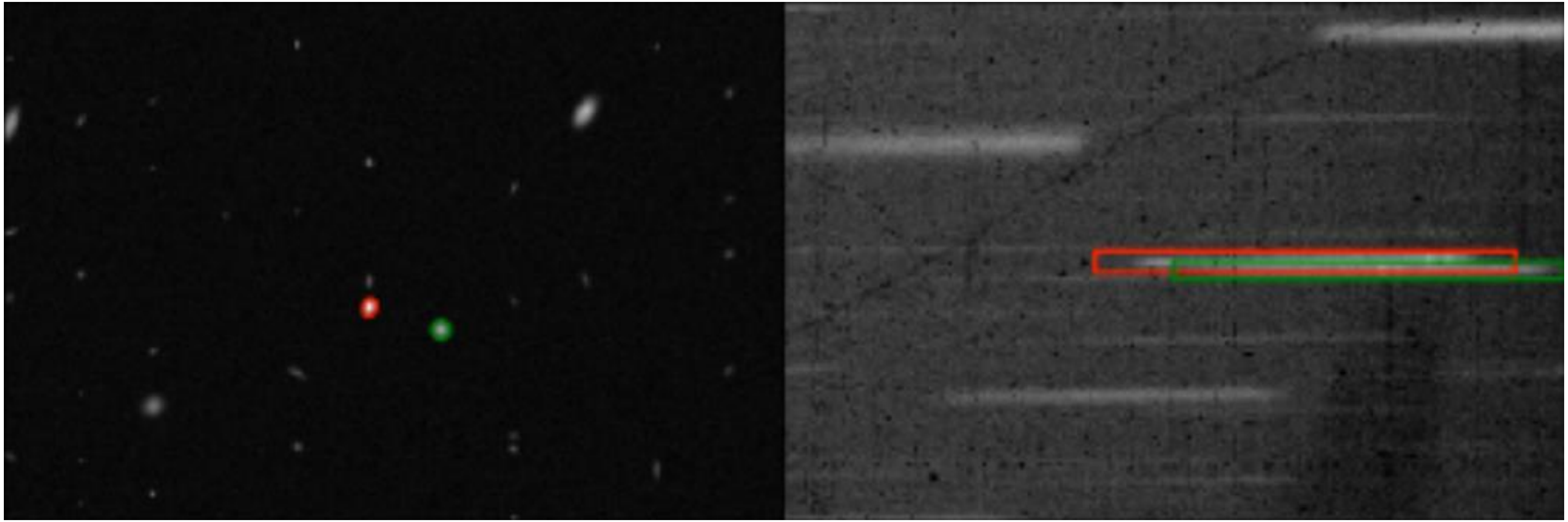


Spectre de la nova Vul avec les raies d’émission d’Hydrogène

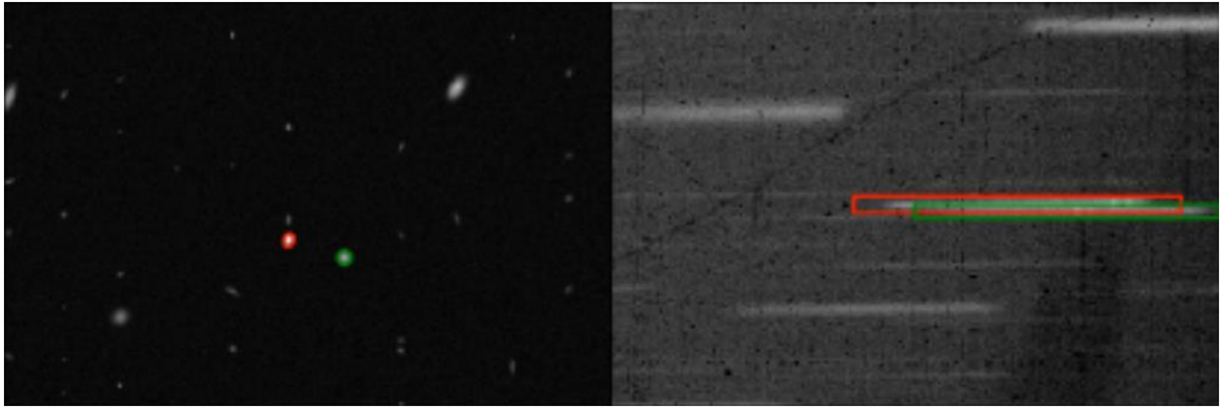
Mélange des spectres

Objets célestes étudiés

Spectres mélangés des objets



Stage de professionnalisation



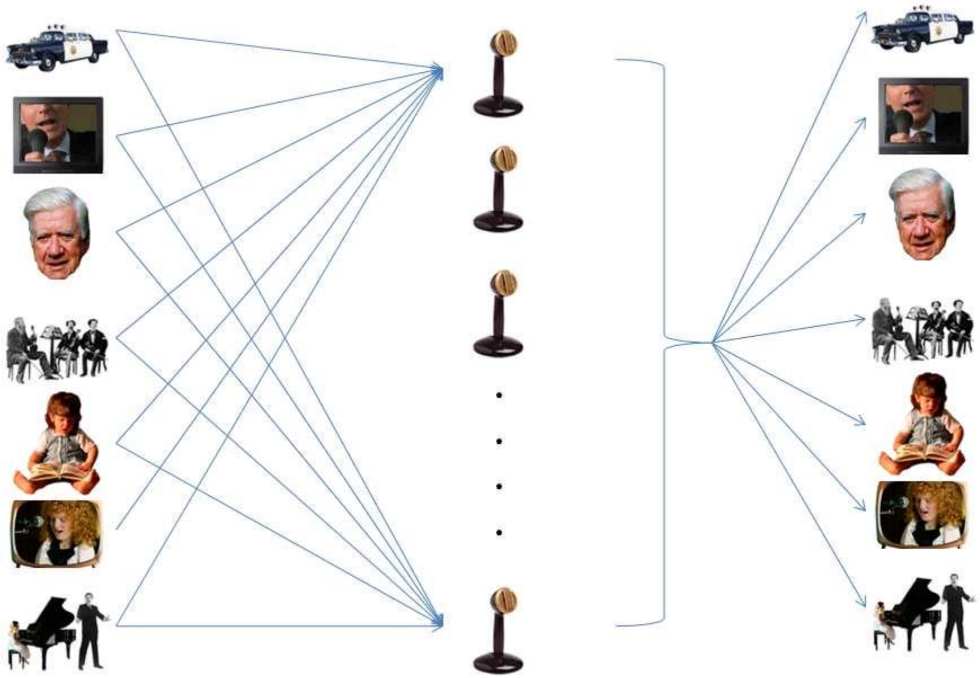
Mettre en oeuvre différents algorithmes de séparation des sources sous MatLab et **les appliquer aux données spectrales** de galaxies.

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

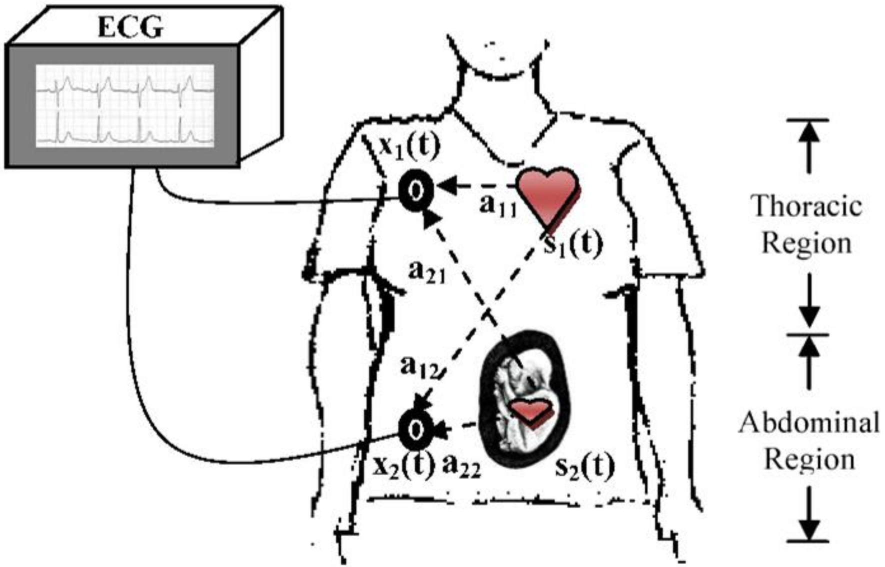
1. Présentation de l'IRAP et CNRS
2. Contexte du stage
3. Sujet de stage - *Séparation des sources*
4. Déroulement de stage
5. Bilan personnel - Conclusion

Séparation des sources ou SAS



Sources *Observations* *Estimations des sources*

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

Analyse en composantes parcimonieuses (SCA)

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)

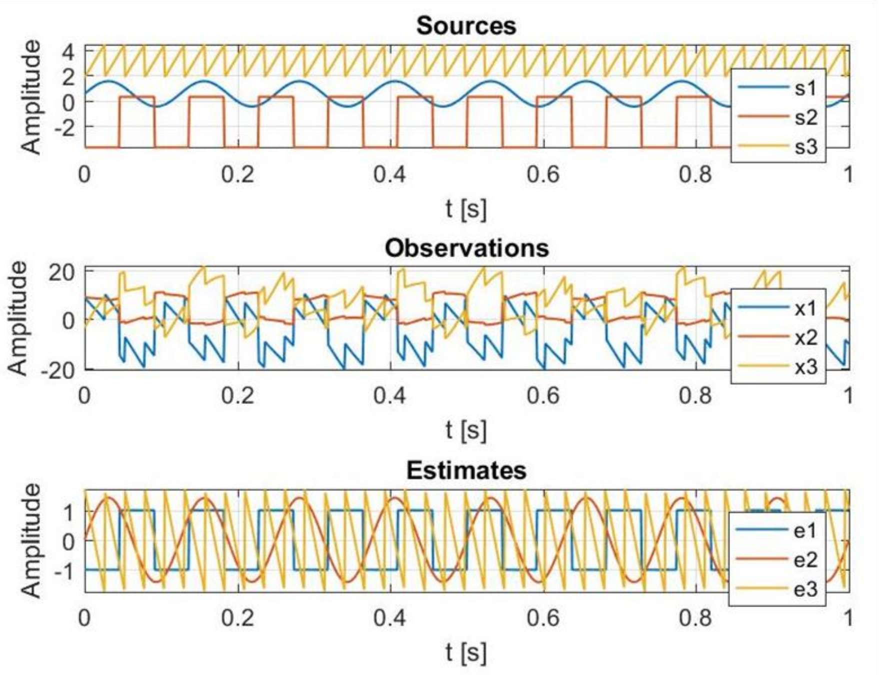
Kurtosis est une mesure de gaussianité. Pour les variables de moyenne nulle il se calcule comme suit $\boldsymbol{kurt(y)= E\{y^4\}-3[E\{y^2\}]^2}$. L’algorithme choisi pour optimiser le critère est l’algorithme du gradient.

```
while 1
    y = w1' * X;
    y3 = y.^3;

    X1 = X';

    grad = 4 *  ( mean([ X1(:,1).*y3', X1(:,2).*y3']) - (3*w1') );

    w1 = w1 + (mu*grad');
    w1 = w1 / norm(w1);

    if (iter==150)
        break;
    end
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

Décomposition en matrices non-négatives (NMF)

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

Décomposition en matrices non-négatives (NMF)

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

Algorithme du gradient

```
x = A*S;

gradA = -(X - x) * S';
A = A - (mu*gradA);
A = max(A,eps);

gradS = -A' * (X-x);
S = S - (mu*gradS);
S = max(S,eps);

e = ((X-A*S).^2)./2;
err = sum(sum(e));
```

Méthode multiplicative

```
A = A.*((X*S')./(A*S*S'));
A = max(A,eps);

S = S.*((A'*X)./(A'*A*S));
S = max(S,eps);

e = ((X-A*S).^2)./2;
err = sum(sum(e));
```

Alternating Least Squares

```
S = inv(A'*A)*A'*X;
S = max(S,eps);

A = X*S'*inv(S*S');
A = max(A,eps);

e = ((X-A*S).^2)./2;
err = sum(sum(e));
```

Décomposition en matrices non-négatives (NMF)

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 upper-right 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 printed sheet at 100 dpi. The noise is due, at least, to the printing process (including the halftoning), to the scanning process and to the non-uniformity in the onion skin paper, especially in its transparency.

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.

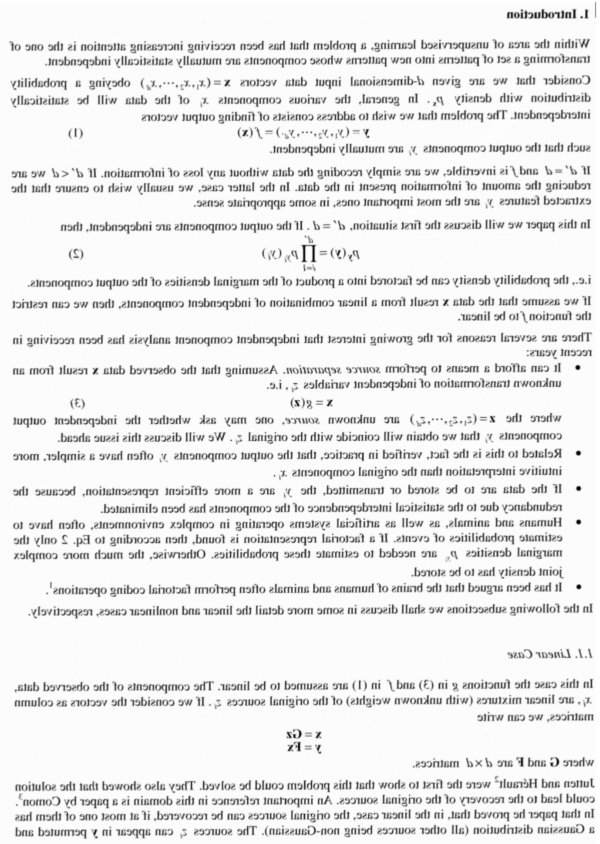
In this example we are creating mixtures that involve natural images, printed text and graphs. The special characteristic of printed text and graphs is that they normally involve just two intensity levels (black and white) although, due to the above mentioned noise, these will appear, in the scanned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

Décomposition en matrices non-négatives (NMF)



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 $\mathbf{x}=(x_1,x_2,\cdots,x_d)$ 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

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

such that the output components y_i 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

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

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 $\mathbf{x}$ result from a linear combination of independent components, then we can restrict the function f to be linear.

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

- It can afford a means to perform *source separation*. Assuming that the observed data $\mathbf{x}$ result from an unknown transformation of independent variables z_i, i.e.

$$\mathbf{x}=g(\mathbf{z}) \quad (3)$$

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

- Related to this is the fact, verified in practice, that the output components y_i often have a simpler, more intuitive interpretation than the original components x_i.
- If the data are to be stored or transmitted, the y_i 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_{y_i} 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[1].

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

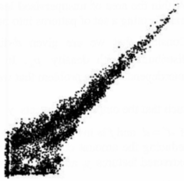
$$\mathbf{x}=\mathbf{G}\mathbf{z}$$
$$\mathbf{y}=\mathbf{F}\mathbf{x}$$

where $\mathbf{G}$ and $\mathbf{F}$ are $d\times d$ matrices.

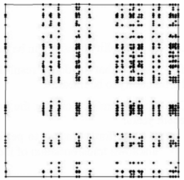
Jutten and Hérault[2] were the first to show that this problem could be solved. They also showed that the solution could lead to the recovery of the original sources. An important reference in this domain is a paper by Comon[3]. In that paper he proved that, in the linear case, the original sources can be recovered, if at most one of them has a Gaussian distribution (all other sources being non-Gaussian). The sources z_i can appear in $\mathbf{y}$ 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 upper-right 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 printed sheet at 100 dpi. The noise is due, at least, to the printing process (including the halftoning), to the scanning process and to the non-uniformity in the onion skin paper, especially in its transparency.



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.

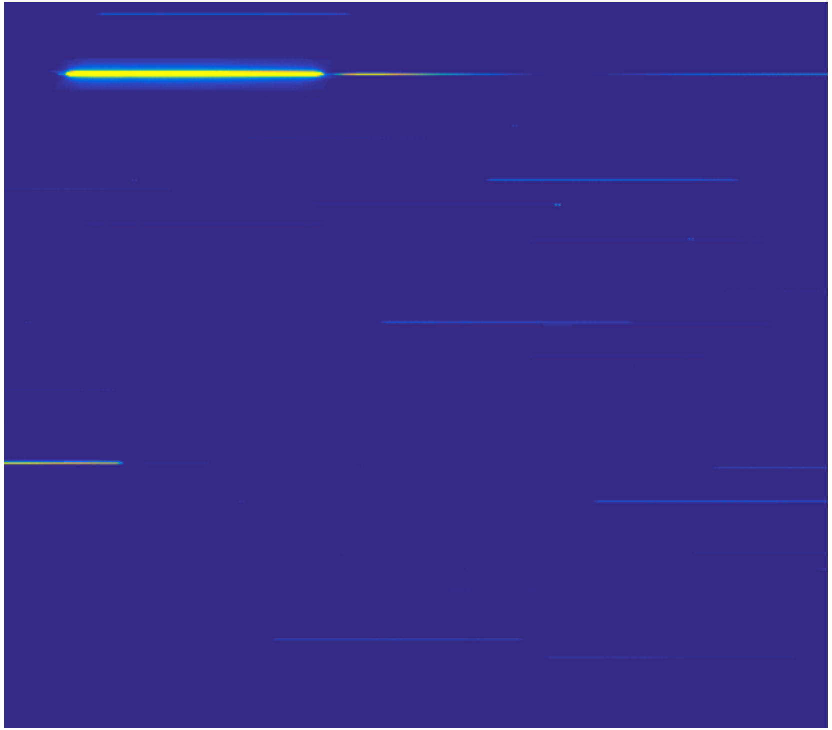
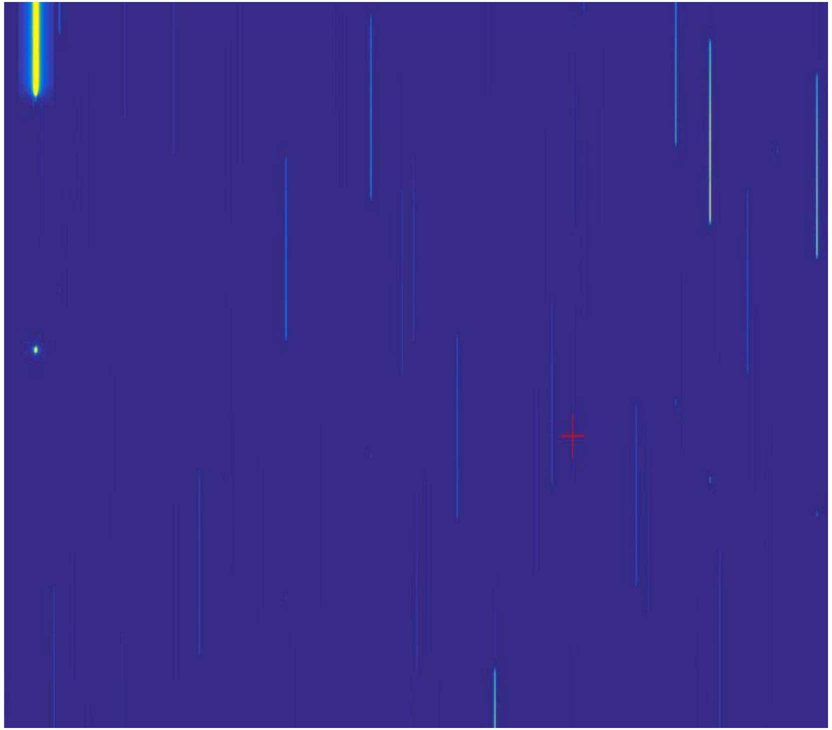
In this example we are creating mixtures that involve natural images, printed text and graphs. The special characteristic of printed text and graphs is that they normally involve just two intensity levels (black and white) although, due to the above mentioned noise, these will appear, in the scanned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

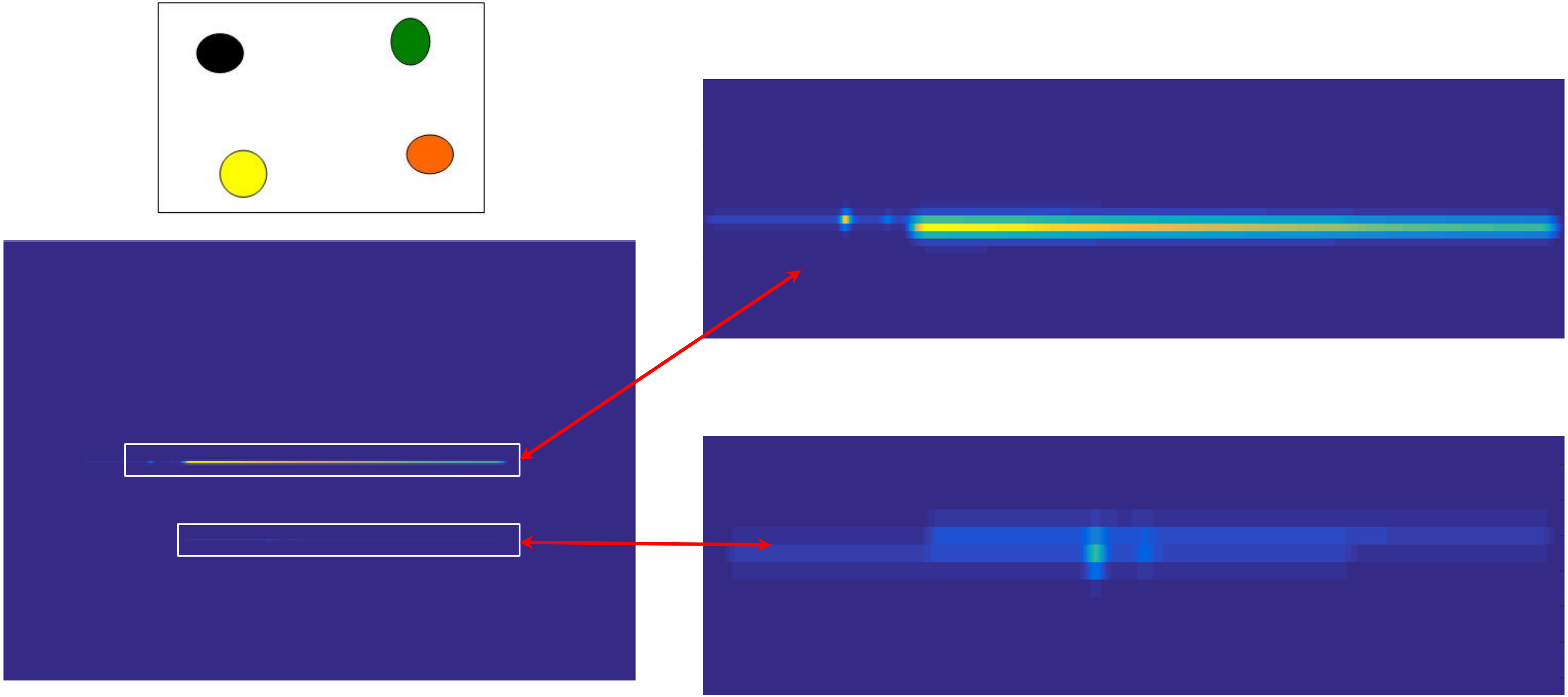
Another issue of interest is the definition of separation criteria that are more suited for images or for printed documents than statistical independence. In fact, images and/or text from the opposite pages of a printed document can easily happen not to be independent from one other. For examples, images 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.

Données spectrales d'EUCLID

Images simulées des spectres des objets célestes

Données spectrales d'EUCLID

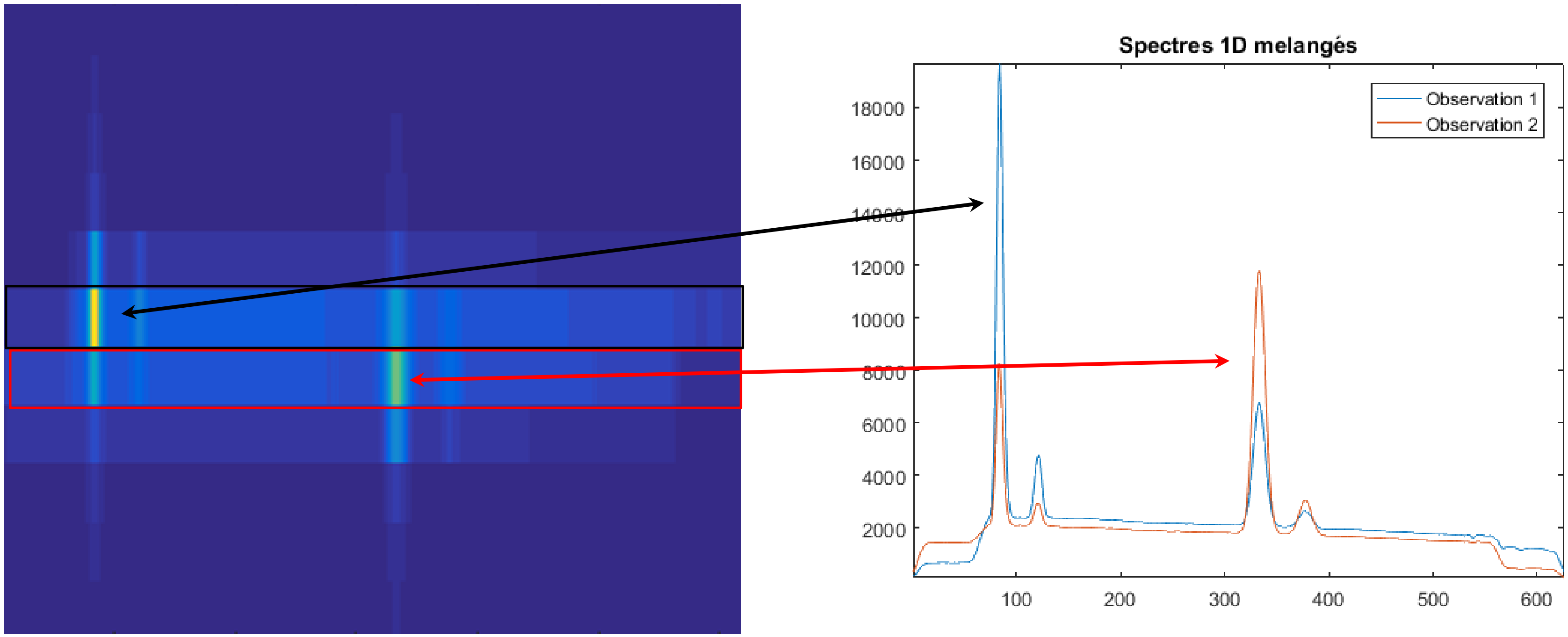


Scénario avec 4 objets

Zoom sur les spectres des objets

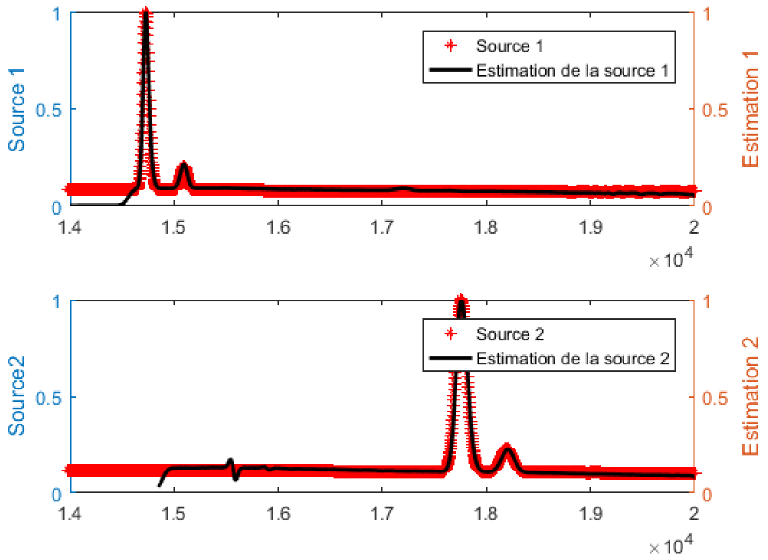
Données spectrales d'EUCLID

1 image - 2 sources mélangées

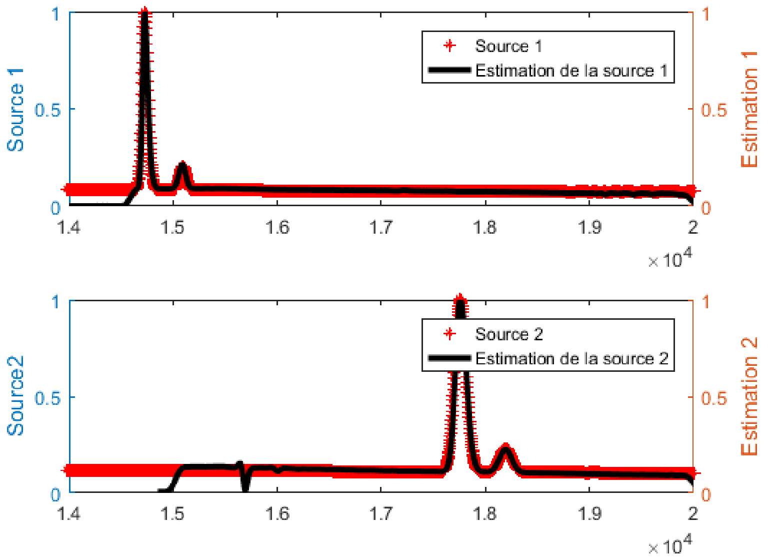


Données spectrales d'EUCLID

1 image - 2 sources mélangées



NMF, méthode multiplicative



ICA, algorithme du gradient

Plan

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

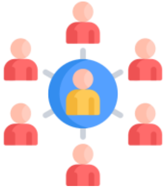
acquises

méthodes de Séparation des sources

notions théoriques en spectroscopie spatiale

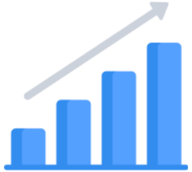
algorithmes d’optimisation

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 time-management



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