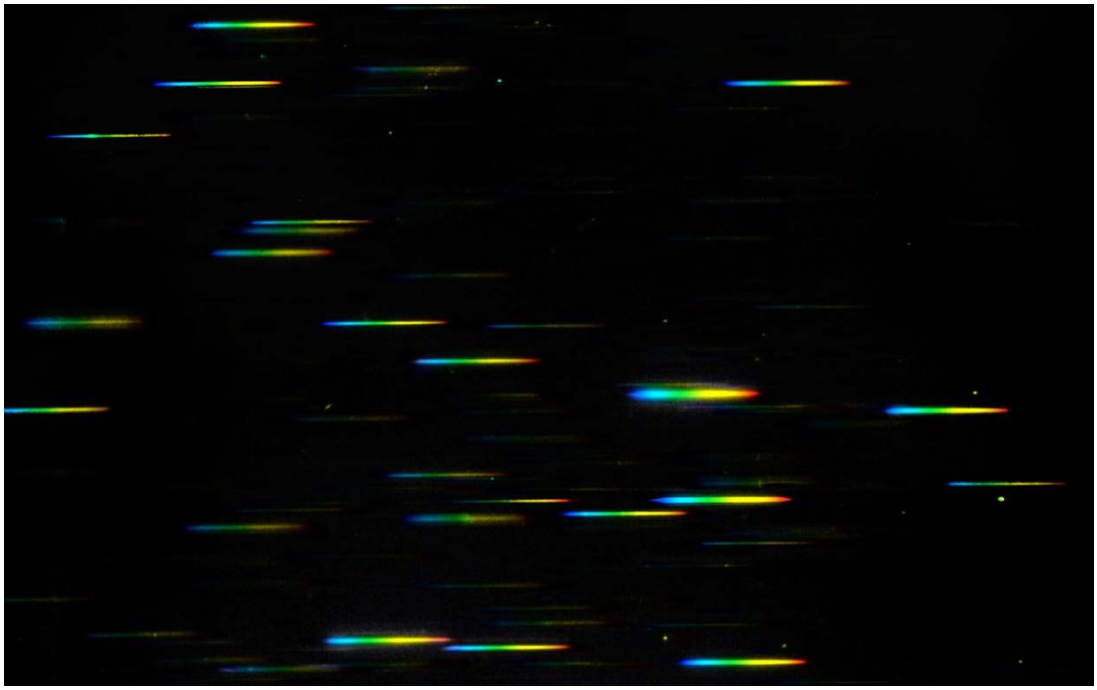
Rapport de stage

Séparation aveugle des spectres de galaxies



Juin - Août 2019

Encadrant de stage : **Shahram HOSSEINI**

Professeur référent : **Philippe CASTELAN**

Stagiaire : **Daria MALIK**

Institut de recherche en astrophysique et planétologie – Signal Images en Science de l'Univers
14, avenue Edouard Belin 31400 TOULOUSE Tél : 33 (0)5 61 33 29 29

Licence 3 EEA CMI– Université Paul Sabatier Toulouse III - 2018/2019

Remerciements

Je tiens à remercier vivement mon maître de stage, M. Shahram HOSSEINI, pour son accueil, le temps qu'il m'a consacré et le partage de son aide et ses conseils au quotidien. Je le remercie également pour m'avoir permis d'effectuer ce stage au sein du laboratoire et m'avoir découvert tout un autre contexte du travail différent de celui d'une entreprise.

Merci aux stagiaires de M1 et M2 et aux doctorants de l'IRAP pour l'aide qu'ils ont pu m'apporter et pour un accueil chaleureux. Je remercie également l'ensemble du groupe Signal Images en Science de l'Univers pour leur bonne humeur et les échanges constructifs.

Résumé

Dans le cadre de ma troisième année de Licence EEA, j'ai réalisé mon stage du 26.06.19 au 17.07.19 et du 12.08.19 au 30.08.18 au sein du groupe Signal Images en Science de l'Univers (SISU) de l'Institut de recherche en astrophysique et planétologie (IRAP).

Ce stage a été l'opportunité pour moi d'enrichir mes connaissances en Traitement du signal et de l'image ainsi que d'appréhender le contexte du travail dans une unité de recherche.

Mon objectif principal était de mettre en œuvre les différents algorithmes de séparation des sources sous MatLab et les tester sur les spectres de galaxies. Il s'agit des données issues des instruments du satellite EUCLID dont le lancement est prévu en Juin 2022. Le satellite étant en conception, les données spectrales utilisées sont celles simulées par un logiciel du CNES qui rapproche au mieux les simulations aux données réelles attendues.

À la fin de ces tests j'ai été amenée à faire une analyse comparative des méthodes mis en œuvre vis-à-vis leurs performances sur les données spectrales et les résultats souhaités et obtenus.

Pour atteindre ces objectifs j'ai divisé mon travail en plusieurs étapes :

- ☑ effectuer une étude bibliographique sur la séparation aveugle des sources, les instruments d'EUCLID et les concepts physiques utilisés dans leur fonctionnement;
- ☑ acquérir des notions mathématiques du domaine de traitement du signal nécessaires pour la mise en œuvre des méthodes de séparation des sources;
- ☑ traduire ces méthodes en algorithmes MatLab et les tester sur plusieurs séries de données différentes (données aléatoires, signaux audio, images) afin de mieux comprendre le concept de la séparation des sources;
- ☑ adapter les scripts MatLab si besoin et les tester sur les données spectrales simulées;
- ☑ faire une analyse comparative des méthodes et conclure sur les performances et la qualité des données obtenues.

Présentation du CNRS et de l'IRAP

Centre National de Recherche Scientifique

CNRS, le Centre National de la Recherche Scientifique, est un établissement public à caractère scientifique et technologique (EPST). Les missions principales du CNRS sont :

- mener les recherches scientifiques présentant un intérêt pour l'avancement de la science ainsi que pour le progrès économique, social et culturel du pays;
- valoriser les résultats et faire bénéficier la société des avancées scientifiques;
- partager les connaissances;
- accueillir des futurs chercheurs, doctorants et post-doctorants pour la recherche;
- contribuer à la politique scientifique et participer à la stratégie nationale de recherche.

L'organisation du CNRS comprend le Directoire, les Directions générales déléguées aux ressources, à la science et à l'innovation, les Délégations régionales, les Instances rattachées à la présidence, les Unités de recherche, le Conseil d'administration, et enfin le Comité nationale de la recherche scientifique.

Le CNRS compte environ 1 100 laboratoires répartis sur l'ensemble du territoire français. Ce sont en très grande majorité des unités mixtes de recherche (UMR) associées à une université, une école supérieure ou un autre organisme de recherche. À ces laboratoires s'ajoutent 36 unités mixtes internationales (UMI).

Les équipes, formées de chercheurs, ingénieurs et techniciens, sont à l'origine de la production et de la transmission des connaissances.

Institut de Recherche en Astrophysique et Planétologie

L'Institut de Recherche en Astrophysique et Planétologie IRAP a été fondé au 1er janvier 2011, à la suite de la fusion de plusieurs laboratoires toulousains. L'IRAP est une Unité mixte de recherche (UMR 5277) du CNRS et de l'Université Paul-Sabatier. L'Institut est localisé à Toulouse et à Tarbes sur les campus de l'Observatoire Midi-Pyrénées. L'unité mixte de recherche est une instance principale dans l'organisation de la recherche en France, disposant de lignes budgétaires propres et de personnel affecté par les partenaires (CNRS ou université, par exemple). Administrée par un directeur et un conseil de laboratoire, elle définit sa stratégie de recherche de manière largement autonome.

Secteur d'activité

Les objectifs scientifiques de l'IRAP concernent l'étude et la compréhension de l'Univers et de son contenu : la Terre en tant que planète, son environnement spatial, le soleil et ses planètes, les étoiles et leurs systèmes planétaires, le milieu interstellaire, les galaxies, les tous premiers astres et le Big Bang primordial.

Une autre partie importante de l'activité du laboratoire est consacrée à la proposition et au développement de projets instrumentaux au sol et dans l'espace. Ces projets s'inscrivent dans le cadre de missions gérées :

- par des agences internationales telles que l'ESO (European Southern Observatory) et le CFHT (Canada-France-Hawaii Telescope) pour le segment Sol;
- par les agences spatiales françaises (CNES), européenne (ESA), américaine (NASA) mais aussi japonaises (JAXA), chinoises et indiennes pour le segment Espace. Les développements instrumentaux sont dans tous les cas financés par l'ESO, le CNES (hardware), le CNRS et l'UPS (salaires).

Les projets sont essentiellement menés sous l'égide des organismes nationaux ou internationaux cités ci-dessus et en collaboration avec d'autres laboratoires de recherche français ou étrangers. Ces développements instrumentaux s'inscrivent dans un planning défini au départ. Cependant, les aléas de la mise au point de ces systèmes complexes imposent aux personnels de s'adapter à ces périodes de suractivité. Ces impératifs ont conduit à une organisation par projet.

Organisation et dimension de l'IRAP

Comprenant environ 300 personnels et étudiants, l'IRAP est un des pôles majeur de l'astrophysique sol-espace en France. Il comprend un effectif de personnels techniques qualifiés dans le domaine de la conception, construction, intégration et exploitation d'instruments au sol et dans l'espace. Il comprend également les experts exerçant les expériences de laboratoire permettant de caractériser les processus physiques.

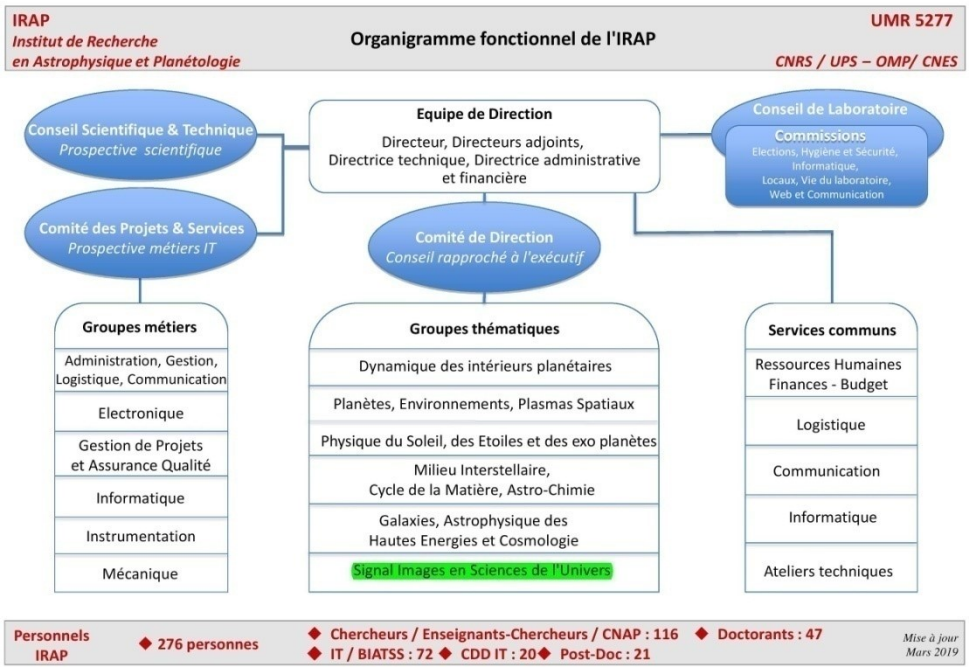


Figure 1 : Organigramme de l'IRAP © IRAP

L'équipe de direction est composée actuellement:

- du directeur : Philippe Louarn
- des directeurs adjoints : Emmanuel Caux, François Lignières et Patrick Pinet
- d'une directrice technique : Laurence Lavergne
- d'une directrice administrative et financière : Nicole Le Gal

L'équipe de direction se réunit tous les 15 jours. À cette réunion sont également invités selon l'ordre du jour la responsable du pôle RH, le responsable du pôle financier, le responsable du pôle logistique et le responsable du service informatique.

Le Comité de Direction (équipe de direction, responsables et responsables-adjoints des groupes thématiques, et conseillers spéciaux de la direction) se réunit également tous les 15 jours afin de se donner les nouvelles sur l'avancement des travaux et les problématiques éventuelles.

Plusieurs conseils et pôles opérationnels effectuent leurs missions au sein de l'IRAP :

- le **Conseil Scientifique et Technique** est en charge de la stratégie scientifique de l'IRAP. Il a un rôle consultatif et discute, délibère et émet des recommandations sur toute question qui impacte sur la stratégie scientifique de l'IRAP à court, moyen et long termes (engagement des projets, priorités sur les sujets de thèse, actions scientifiques de l'IRAP);
- le **Comité de Projets et Services** est en charge de l'animation et de la prospective des métiers ITAs[1] à l'IRAP. Sa composition comprend le directeur, la directrice technique, la Directrice Administrative et Financière, un directeur adjoint, les responsables de groupes techniques et la responsable du pôle des ressources humaines. Le Comité se réunit avec les chefs des projets et des services concernés lorsqu'il doit discuter le plan de charge des ITAs des Groupes métiers;
- le **Conseil de Laboratoire** est l'instance statutaire d'administration de l'IRAP. Il émet des avis sur toutes les questions relatives à la politique scientifique, la gestion des ressources, l'organisation et le

[1] Ingénieurs, techniciens et personnels administratifs

fonctionnement de l'IRAP. Il conduit certaines de ces actions au moyen de Commissions qu'il nomme et qu'il pilote: "Vie du laboratoire", "Informatique", "WEB COM", "Locaux", "Élections", "Hygiène et sécurité", "Formation permanente". Les directeurs adjoints sont invités permanents des réunions, et les responsables des groupes thématiques sont invités en fonction de l'ordre du jour;

- les **Services communs** réunissent plusieurs pôles fonctionnels tels que Ressources Humaines, Finances, Logistique etc.

Les six groupes thématiques de l'IRAP se composent de plusieurs chercheurs, enseignants-chercheurs, doctorants et post-doctorants :

- Galaxies, Astrophysique des Hautes Énergies et Cosmologie (GAHEC)
- Planètes, Environnements et Plasmas Spatiaux (PEPS)
- Dynamique des Intérieurs Planétaires(DIP)
- Milieu Interstellaire, Cycle de la Matière, Astro-Chimie (MICMAC)
- Physique du Soleil, des Étoiles et des Exoplanètes (PS2E)
- Signal-Images en Sciences de l'Univers (SISU)

Ces équipes scientifiques mènent des recherches fondamentales afin de répondre aux questions d'actualité et faire progresser la communauté scientifique.

En ce qui me concerne, j'ai effectué mon stage au sein du groupe thématique SISU - Signal Images en Science de l'Univers (Fig. 1). SISU regroupe des astronomes, ingénieurs et chercheurs autour des thèmes de recherche liés au traitement du signal et des images en Sciences de l'Univers. Le groupe bénéficie d'étroites collaborations avec les autres équipes du laboratoire ou de l'Observatoire Midi-Pyrénées. Les thèmes forts étudiés se répartissent ainsi sur la formation et reconstruction des images, modélisation et analyse des signaux et images et séparation aveugle de sources.

L'équipe s'investit également dans les projets instrumentaux. Ainsi, par exemple, le laboratoire est reconnu par ESA en tant que Expert Support Laboratory sur les questions de reconstruction d'images et de calibration de l'instrument de la mission SMOS (Soil Moisture and Ocean Salinity) pour l'observation de la Terre. Les chercheurs de SISU sont également impliqués dans le traitement de données pour le projet de spectro-imageur MUSE issu d'un consortium de 7 laboratoires européens (dont l'IRAP). En partenariat avec 3 autres laboratoires français (CRAL-Lyon, OCA-Nice, LSIIT-Strasbourg), l'équipe fait partie du projet ANR DAHLIA sur le traitement de données hyperspectrales en général, et MUSE en particulier.

Contexte du projet de stage

Comme il a été évoqué plus haut, mon travail portait sur les données spectrales issues d'un logiciel simulateur du CNES. Le satellite aura pour mission de mesurer et collecter les spectres de plusieurs millions de galaxies. Ces données seront ensuite analysées par des scientifiques afin d'estimer les décalages spectraux (redshift) des galaxies et comprendre plus sur l'expansion de notre Univers et l'énergie noire.

En premier lieu j'ai dû me documenter sur EUCLID et d'autres questions concernant la problématique de mon sujet de stage. Dans ce qui suit je vais résumer une partie de mes études bibliographiques concernant la notion de décalage spectral, le fonctionnement des instruments d'EUCLID, ainsi que les principes de séparation aveugle des sources. Cette étape était indispensable pour moi pour appréhender mieux le contexte du sujet de stage et acquérir les connaissances nécessaires à la réalisation des objectifs définis.

Spectroscopie - Décalage spectrale

Avant d'aborder la notion de redshift il est indispensable de comprendre ce qui est la spectroscopie en général.

La spectroscopie, ou spectrométrie, est l'ensemble des techniques permettant d'analyser la lumière émise ou réfléchie par un corps étudié. En effet, l'étude expérimentale du spectre électromagnétique d'un phénomène physique nous permet de dévoiler sa décomposition sur une échelle d'énergie, ou toute autre grandeur se ramenant à une énergie (fréquence ou longueur d'onde, par exemple ; Figure 2).

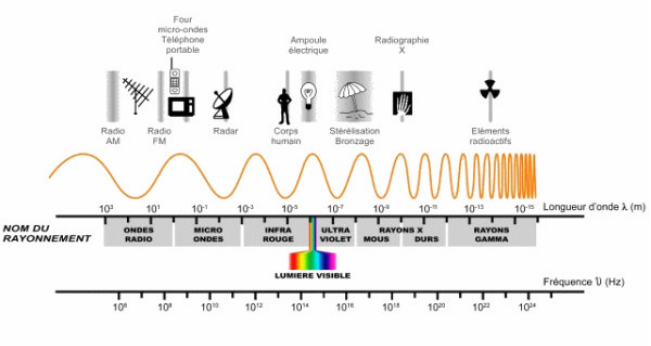


Figure 2 : Spectre électromagnétique © atala.fr

Tout corps réel peut émettre, absorber ou réfléchir des ondes électromagnétiques qu'il s'agisse de l'ultraviolet, visible, infrarouge etc. De ce fait on distingue deux types de spectres.

Spectre d'absorption : si le corps est exposé à un rayonnement, l'énergie portée par les ondes peut exciter les électrons des atomes du corps. Si la quantité d'énergie est assez importante, les électrons passent à leurs autres niveaux énergétiques. Les photons, absorbés par les électrons excités, ne sont plus présents dans le flux du rayonnement. En faisant passer la lumière traversant le corps dans un prisme, nous pouvons la disperser et observer les raies noires sur le fond coloré (Figure 3). Le fond coloré - l'arc-en-ciel - est la lumière blanche dispersée, et les raies noires correspondent aux longueurs d'ondes des photons absorbés. S'il s'agit d'un rayonnement non visible, nous allons observer les piques vers le bas signifiant la quantité très basse des photons reçus. Ce qui veut dire que ces photons ont été absorbés (Figure 4).

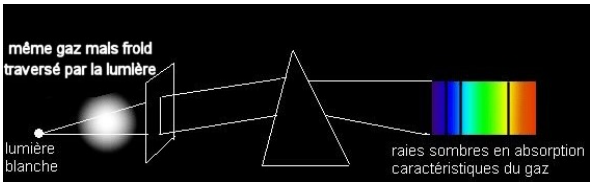


Figure 3 : Spectre d'absorption d'un gaz froid éclairé par la lumière blanche © Wikipedia

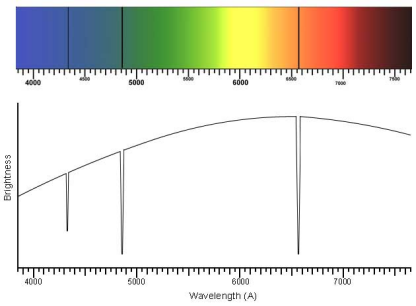


Figure 4 : Spectre d'absorption de l'hydrogène © ia.terc.edu

Spectre d'émission : il s'agit des corps suffisamment chaud pour émettre eux-mêmes un rayonnement dû à sa température, ou des corps bombardés par des électrons extérieurs. Ces électrons extérieurs apportent de l'énergie aux atomes du corps. Les électrons excités changent de niveaux énergétiques et évacuent de l'énergie sous forme de rayonnement. Parmi les spectres d'émission il existe des spectres continus et des

spectres de raies d'émission. Les spectres continus - un arc-en-ciel complet - correspond à la lumière blanche ou aux corps très chauds et très dense, comme les étoiles ou les galaxies (Fig. 5). Les raies d'émission correspondent, quant à elles, aux corps simples ou aux mélanges des atomes indépendants qui n'interagissent pas entre eux (Fig. 6). Nous observons les raies de couleur sur un fond noir pour le rayonnement visible (Fig. 6,7). Ces raies correspondent aux photons émis par le corps. Pour le non-visible, les piques vers le haut signifient que le capteur reçoit les photons des longueurs d'ondes correspondantes en grande quantité (Fig. 6). Ce sont donc les photons émis par le corps.

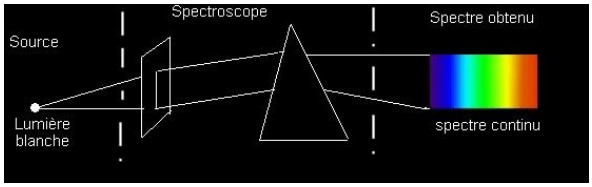


Figure 5 : Spectre continu de la lumière blanche © Wikipedia

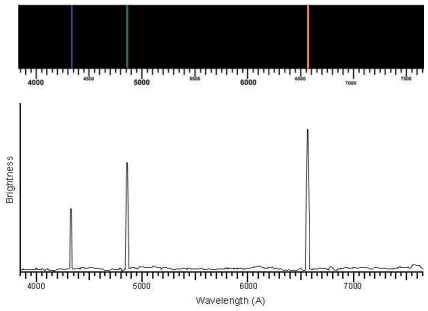


Figure 6 : Spectre d'émission de l'hydrogène © ia.terc.edu

Chaque élément chimique possède ainsi des raies d'émission/absorption qui lui sont propre comme une empreinte digitale (Fig. 7). En se référant à ces "empreintes digitales" connues, il est possible de déduire la composition d'un corps chimique ou d'une étoile, ainsi que de détecter la présence des éléments non-désirable s'il s'agit de l'analyse des aliments, par exemple.

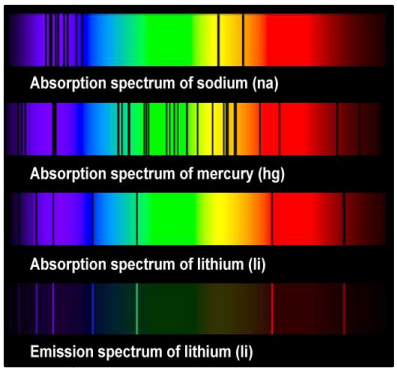


Figure 7 : Spectres de certains éléments chimiques © astronoo.com

Actuellement les scientifiques dans les domaines d'astronomie et cosmologie possèdent énormément d'informations sur plusieurs milliers d'étoiles, de galaxies et d'autres corps célestes. Cependant, en observant les spectres des galaxies ou étoiles, les scientifiques remarquent un décalage de l'ensemble de spectre de ces objets à droite, vers le rouge - *redshift* (Fig. 8) - par rapport aux références de laboratoire.

Ce phénomène est connu sous le nom de décalage spectral vers le rouge. Cependant, la lumière émise ou reçue n'est pas nécessairement rouge. Le terme "redshift" fait référence à la perception de l'œil humain des ondes plus longues comme le rouge. Dans le sens physique le "redshift" signifie une *augmentation* de la longueur *de n'importe quelle onde* du rayonnement électromagnétique. Cela peut également être les rayons gamma perçus en tant que les rayons-X, ou la lumière visible perçue en tant que les ondes infra-rouge ou les ondes radio (Fig. 9).

Les ondes peuvent également subir une déformation dans le sens opposé - *diminution* de la longueur d'onde. Dans ce cas les raies du spectre sont décalées à gauche, dans le sens décroissant de la longueur d'onde. Nous parlons de décalage vers le bleu, ou *"blueshift"* en anglais.

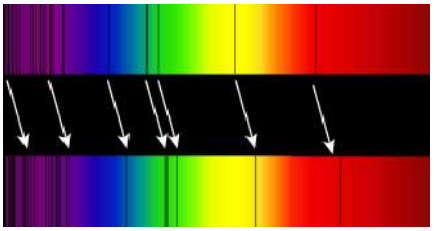


Figure 8 : Le spectre visible du Soleil (en haut) comparé à un spectre d'un groupe de galaxies lointaines (en bas) © Wikipedia

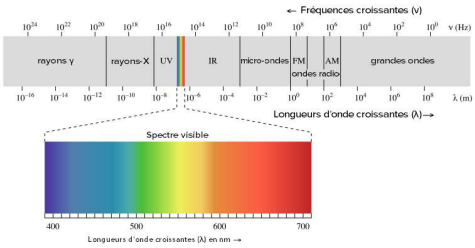


Figure 9 : Spectre électromagnétique © fr.khanacademy.org

Dans les domaines d'astronomie et cosmologie les scientifiques mettent en évidence une cause principale de ces déformations d'ondes - l'expansion de l'Univers. L'expansion de l'Univers est le phénomène qui voit à grande échelle les objets composant l'Univers s'éloigner les uns des autres. Les ondes sont étirées suite à un gonflement de l'espace lui-même (Fig. 10,11). Cependant, l'expansion n'affecte pas la taille des objets à cause de leur gravité "intérieur", et les objets ne changent pas d'endroits car ils ne bougent pas (Fig. 10).

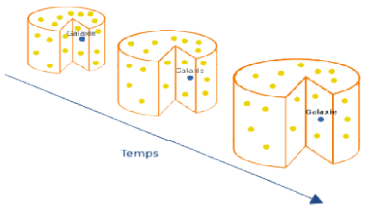


Figure 10 : L'expansion de l'Univers imagée par le gonflement d'un gâteau © Wikipedia

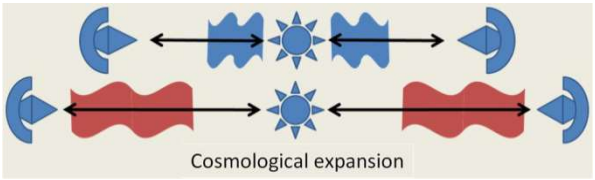


Figure 11: Elongation d'ondes dû à la dilatation de l'espace et l'expansion de l'Univers© Wikipedia

Le décalage spectral peut être également provoqué par le mouvement des objets célestes, le résultat de l'effet Doppler (Annexe, *L'effet Doppler*). C'est une cause pourtant très rare dont les effets sont beaucoup moins important devant ceux de l'expansion de l'Univers. Le seul exemple de décalage spectral dû au mouvement d'un objet est la galaxie d'Andromède, qui est une galaxie voisine de la nôtre et dont le spectre est décalé vers le bleu, ou *"blueshifted"*, car la vitesse à laquelle elle *s'approche* de nous est plus importante que l'effet de l'expansion de l'Univers.

En ce qui concerne les galaxies qui vont être photographiées par EUCLID, les scientifiques possèdent déjà certaines informations sur leurs structures et positions. Ils sont donc capables d'obtenir les références de laboratoire et les comparer avec les données spectrales du satellite afin d'estimer le redshift.

Les instruments d'EUCLID - Mélanges des spectres

Afin de répondre aux besoins des scientifiques, EUCLID sera équipé de deux instruments principaux pour sa mission : un instrument Imageur Visible (VIS) qui fournira des images de haute qualité dans le spectre visible et un instrument pour le rayonnement proche-infrarouge (NISP).

L'instrument NISP est composé:

- d'un spectromètre pour étudier les spectres ;
- d'un photomètre pour étudier le rayonnement lumineux des objets célestes.

Pour effectuer la spectro- et photométrie le NISP sera équipé de deux mécanismes à roues.

Une des deux roues contient des filtres ("Filter Wheel" sur la Fig. 12) pour la photométrie. Cette chaîne fera les images du ciel dans les 3 bandes de longueurs d'ondes proche infra-rouge (Fig. 13). Pour la photométrie la roue à filtres est en position du filtre nécessaire, tandis que l'autre roue est en position ouverte.

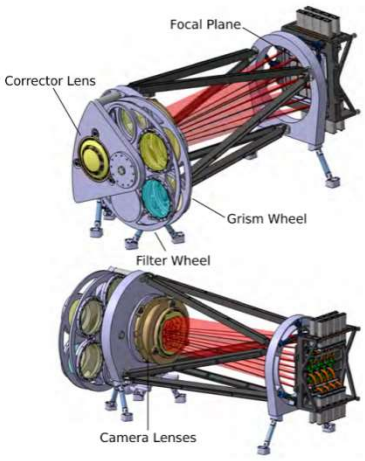


Figure 12 : Instrument NISP d'EUCLID © CNES

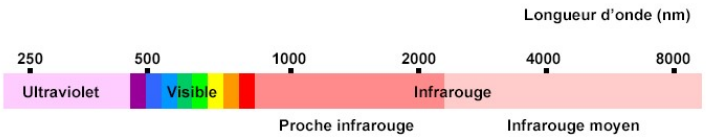


Figure 13 : Spectre électromagnétique proche du visible © geoportail.gouv.fr

La seconde roue est une roue à grismes ("Grism wheel" sur le Fig. 12). Cette roue contient 4 grismes différents et une position ouverte. *Grisme* est une combinaison d'un prisme et d'un réseau de diffraction ("grating" en anglais, *GRating+prISM*). Le réseau de diffraction décompose la lumière incidente sous différents

angles, selon ses longueurs d'onde (ou couleurs) constitutives (Fig. 14). Le prisme permet de rediriger les rayons dans le reste de la caméra afin que nous puissions les voir (Fig. 15).

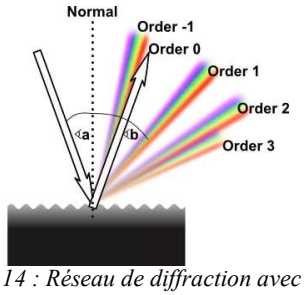


Figure 14 : Réseau de diffraction avec une lumière incidente © quora.com

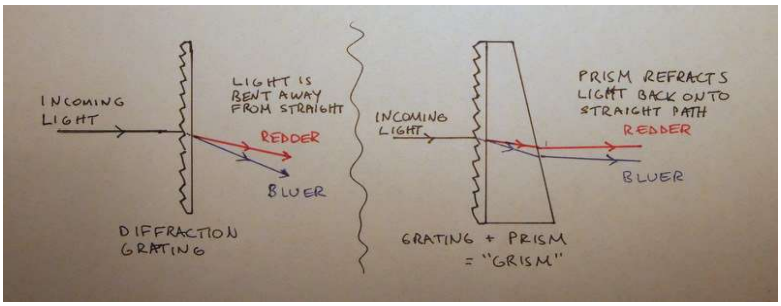


Figure 15 : Différence entre un réseau de diffraction et grisme © candels-collaboration.blogspot.com

Les 4 grismes du NISP vont, de la même façon, couvrir les longueurs d'onde proche infra-rouge. Pour la spectroscopie la roue à filtres est en position ouverte, et la roue à grisme est tourné jusqu'au grisme nécessaire. Derrière les filtres se trouve la caméra qui focalise la lumière sur le plan focal.

L'image d'une galaxie à travers un filtre normal ressemble à une photo classique (Fig. 16, image de gauche) mais l'image faite à travers un grisme est ainsi un spectre étalé, tout comme un arc-en-ciel pour une lumière blanche visible dispersée (Fig. 16, image de droite). Malgré leur continuité, les spectres présentent également des endroits plus brillants que d'autres, ce qui correspond la plupart du temps aux fortes raies d'émission (Fig. 10, le spectre entouré en blanc avec les endroits brillants signés *Hbeta, Halpha* et *He I*).

Figure 16 : Photo du Hubble des galaxies lointaines © ESA/Hubble/NASA

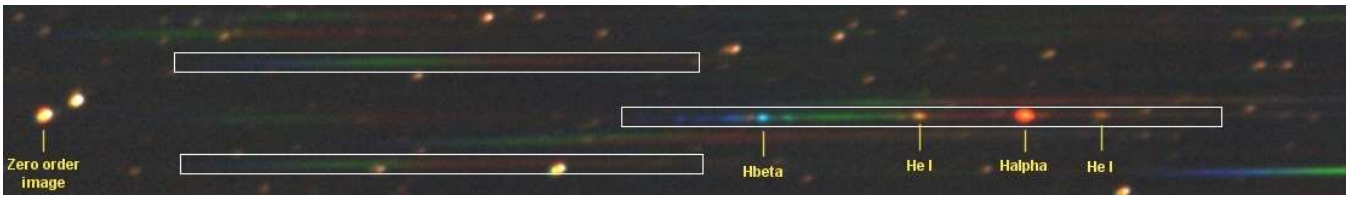


Figure 17 : Spectre de la nova Vul avec les raies brillantes d'hydrogène, 2007 (V458 Vul) © astrosurf.com

Étant donné que le grisme est opposé à la direction de la lumière tout comme un filtre pour la photométrie, l'image réalisée avec un grisme présente des spectres dispersés pour tout ce qui se trouve dans le champ de vision. C'est à la fois une force et une faiblesse.

Cela signifie qu'une image est capable de nous fournir les spectres de nombreux objets à la fois. Mais si deux ou plusieurs corps célestes se trouvent très proches l'un de l'autre ? En effet, leurs lumières étalées vont se chevaucher et les spectres de ces corps seront mélangés (exemples entourés en couleur sur les Fig. 18 et 19).

Les scientifiques ne sont donc plus capables de traiter correctement les données spectrales. Ce problème est également vu comme un problème de *séparation des sources.* Les spectres originaux des corps célestes sont contaminés, il sera utile de les séparer afin de pouvoir estimer le redshift et obtenir le maximum d'informations possibles à partir des données d'EUCLID.

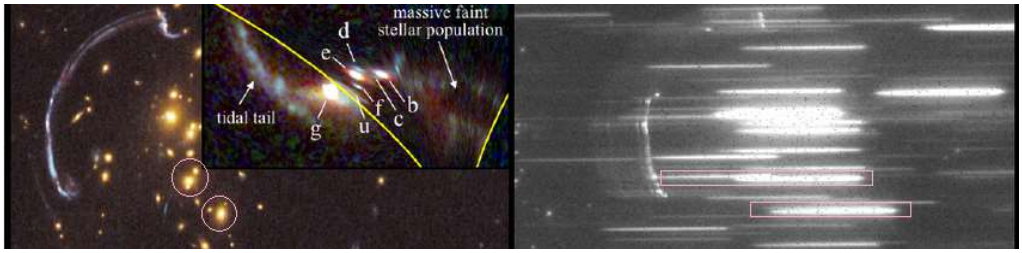


Figure 18 : Spectre de la formation de galaxie RCS0327 © Katherine E. Whitaker, Jane R. Rigby, NASA, 2014

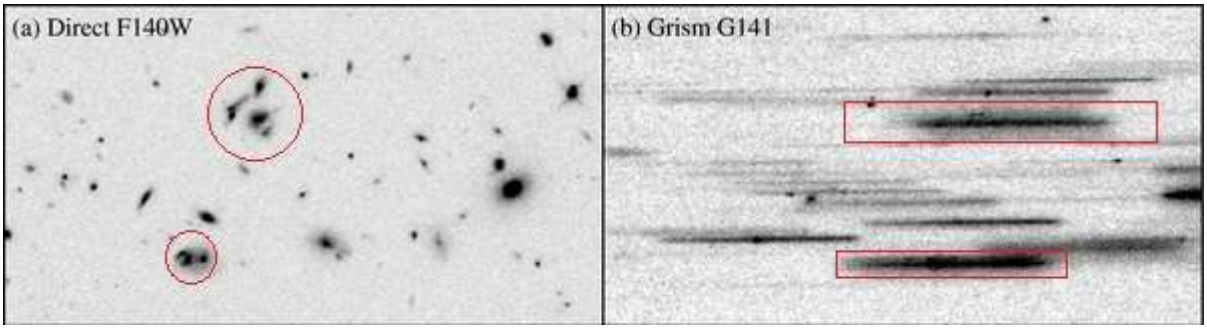


Figure 19 : Quelques spectres obtenus avec le grisme objectif du Hubble © Gabriel B. Brammer et al., The American Astronomical Society, 2012

Séparation des sources

Dans ce qui suit je présenterai un aperçu général du problème de séparation des sources (SAS) et des méthodes classiques de SAS existantes.

Mon travail réalisé porte sur les méthodes de séparation *aveugle* des sources. Cela veut dire que les observations sont connus (les spectres mesurés), et nous n'avons que certaines ou, dans le pire des cas, aucune information sur les sources (les spectres individuels de chacun des corps célestes étudiés). Plusieurs méthodes et techniques existent déjà et fournissent des résultats satisfaisants. Néanmoins, la quantité et la particularité des données que le satellite fournira nécessitent de nouveaux algorithmes plus performants et innovants. Dans cette partie je voudrais présenter quelques notions générales du problème de séparation des sources ainsi que les contraintes auxquelles les scientifiques font face en travaillant avec les données spectrales des galaxies.

Un des exemples classiques qui permet de mieux comprendre le concept de la séparation des sources est un exemple de séparation des signaux audio (Fig. 20). Les signaux des parties des différents instruments sont *les sources S*, les signaux enregistrés issus des microphones sont *les observations X*. Chacune des observations présente un mélange entre les parties de chaque instrument. Nous ne connaissons pas les coefficients de mélange et les contributions des sources dans chacune des observations. Nous cherchons à démêler les observations et en extraire *les estimations* des signaux de source. Il s'agit des estimations car il n'est possible d'estimer les signaux de source qu'à quelques indéterminations près.

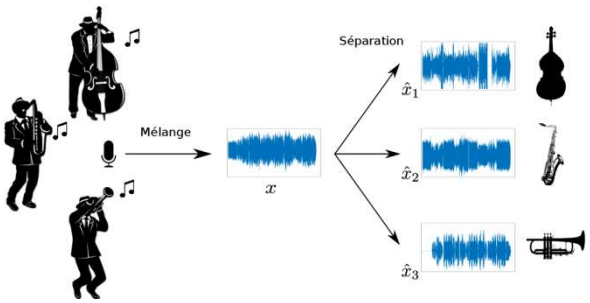


Figure 20 : Exemple d'un mélange des sources audio © semanticscholar.org

Les domaines d'application de SAS sont très variés : séparation des signaux audio et des signaux biomédicaux avec plusieurs capteurs (Fig. 21). Les méthodes de SAS permettent également de traiter les documents scannés. Notamment, il est possible de séparer les côtés recto-verso des documents, quand la qualité des supports en papier (papier trop ancien et gras ou papier semi-transparent) fait que les deux côtés d'un document peuvent être simultanément visibles sur les images (Fig. 22).

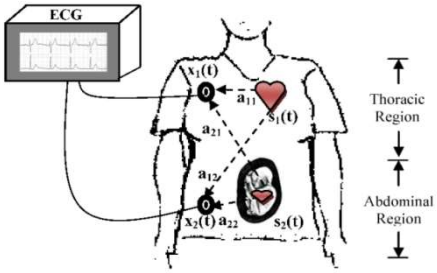


Figure 21 : Exemple d'un mélange des signaux médicaux © semanticscholar.org

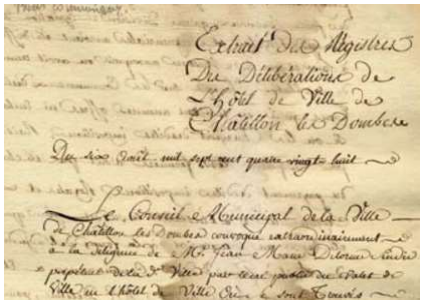


Figure 22 : Exemple d'un mélange des côtés recto-verso d'un ancien document © researchgate.net

Comme il a été mentionné ci-dessus, de nombreuses différentes techniques et méthodes existent pour répondre à la problématique de séparation des sources. Leur validité et la qualité des résultats obtenus dépendent malgré tout de plusieurs critères et caractéristiques des mélanges.

En premier lieu, nous distinguons trois grandes familles des mélanges en fonction de nombre de sources et d'observations :

- mélanges sur-déterminés : nombre d'observations > nombre de sources ;
- mélanges déterminés : nombre d'observations = nombre de sources ;
- mélanges sous-déterminés : nombre d'observations < nombre de sources.

Bien évidemment, les mélanges sur-déterminés et déterminés peuvent être traités avec n'importe quelle méthode et avoir des résultats assez "propres" et satisfaisants. Les mélanges sous-déterminés en revanche sont plus compliqués à manipuler à défaut d'informations fournies par les observations.

Ensuite, nous distinguons différentes *natures* de mélanges en fonction des critères suivants :

- mélange linéaire, par exemple $x(t) = a\,.\,s_1(t) - b.\,s_2(t)$ avec a et b des coefficients scalaires et constants, ou non-linéaire, $x(t) = a\,.\,s_1(t) + b.\,s_1(t).\,s_2(t)$;
- mélange sans bruit ou mélange bruité ;
- mélange sans mémoire (instantané) ou avec mémoire (convolutif), par exemple la présence des échos lors d'un enregistrement audio.

Tous ces critères introduisent des extensions dans les méthodes classiques de SAS qui, elles, se basent sur les différentes natures et caractéristiques des *sources.* Ainsi, il existe trois principales familles de solutions de SAS, listées ci-dessous, correspondantes à trois hypothèses différentes faites sur les sources.

L'Analyse en composant indépendantes, ou ICA (Independent Component Analysis), qui nécessite que les sources soient indépendantes *statistiquement*.

L'Analyse en composant parcimonieuses, ou Sparse Component Analysis, qui évoque la parcimonie des sources. Une source est dite parcimonieuse quand elle a des valeurs nulles dans un domaine de représentation. Par exemple, le signal de parole peut être parcimonieux en raison d'existence des zones de silence.

La **Factorisation en matrices non-négatives**, ou NMF (Non-negative matrix factorization), fait appel aux sources et mélanges qui physiquement ne peuvent prendre que des valeurs positives (non-négatives).

En ce qui me concerne, nous avons conclu avec mon maître de stage que je ne travaillerai que sur deux des trois méthodes de SAS, eu égard aux types de données étudiées : le ICA pour mieux comprendre les principes et effectuer des tests sur les données plus simples, comme les signaux audio et les images recto-verso, et la NMF, car les données spectrales sont toujours non-négatives et actuellement c'est *la* méthode que l'équipe SISU cherche à améliorer pour le futur traitement des données d'EUCLID. Quant à l'analyse en composant parcimonieuses, la plupart des mélanges ne vérifiant pas l'hypothèse de la méthode, celle-ci n'est pas universelle dans notre cas et n'est pas complètement adaptée à ce genre des traitements. Dans ce qui suit je vais présenter ce que j'ai pu accomplir durant mon stage et quelle étaient mes démarches. Je vais également exposer les résultats obtenus et une brève analyse des performances des différents algorithmes que j'ai pu mettre en œuvre.

Etude de problématique et mise en œuvre

Etat de l'art

J'ai commencé mon stage par une étude bibliographique concernant les instruments d'EUCLID et la notion du problème de séparation des sources.

Pour ce faire j'ai du me servir de plusieurs sources d'informations différentes : thèses, articles scientifiques, extraits des conférences, vidéos et sites éducatives, livres.

Presque toutes les sources étaient en format numérique. M. Hosseini en a partagé certaines avec moi au tout début de mon stage. Cela m'a été utile pour "la mise au niveau" et la meilleure compréhension des détails du sujet. À fur et à mesure que j'avançais dans mon travail j'avais besoin de nouvelles informations ou détails que je cherchais cette fois-ci moi-même sur Internet.

Pour les explications générales je me référais aux livres et thèses aussi bien qu'aux sites ou vidéos éducatives pour les étudiants.

En ce qui concerne les calculs et les questions plus mathématiques, je m'adressais aux livres scientifiques et thèses qui présentaient des méthodes et algorithmes différents de solution de SAS que j'allais mettre en œuvre ultérieurement.

Pour tout côté "codage" j'essayais de chercher les solutions moi-même en appliquant les compétences acquises en L3. Dans le cas contraire je visitais le forum et le "manuel" en ligne de MatLab à la recherche des fonctions et outils nécessaires pour mes algorithmes. M. Hosseini m'a également partagé certaines de ces connaissances concernant le logiciel et ses fonctionnalités.

Analyse en composants indépendantes (ICA)

J'ai décidé de commencer par l'étude et mise en œuvre de l'Analyse en composants indépendantes (ICA). Cette famille de solutions est actuellement utilisée pour la plupart des problèmes de séparation des sources. Les méthodes basées sur l'indépendance des sources étaient également les toutes premières méthodes de SAS. Pour le reste de cette partie nous supposons que notre modèle de mélange est linéaire, instantané (sans mémoire), sans bruit et déterminé car nous aurons autant d'observations que de sources.

Comme nous avons vu auparavant, les méthodes d'ICA sont valides et applicables à condition que les sources soient statistiquement indépendantes. En statistique et probabilité les événements sont indépendants s'ils n'ont aucune influence l'un sur l'autre. En d'autres termes, si une information sur un des événements n'apporte aucune information sur les autres. Par exemple, la valeur d'un lancer de dés n'a aucune influence sur un deuxième lancer. Dans le cas d'un enregistrement audio de plusieurs instruments ou d'une musique avec une voix nous pouvons dire que les parties des instruments sont indépendantes entre elles, la voix humaine sera également indépendante de la musique. Pour les documents scannés, les pages recto et verso sont elles aussi indépendantes.

Depuis les premières méthodes des années 80 différentes approches ont été proposées par les scientifiques. Actuellement les méthodes les plus répandues se basent sur la maximisation de la non gaussianité, l'exploitation de la structure des signaux, les mesures de l'information mutuelle et l'estimation de vraisemblance [1]. J'ai choisi de travailler sur la non gaussianité car le concept des méthodes correspondantes m'a paru plus intuitif et universel.

La notion de non gaussianité découle du théorème central limite. Ce théorème dit que la distribution, ou la fréquence d'apparition des valeurs, de la somme des variables aléatoires indépendantes tend toujours vers une distribution gaussienne. La figure ci-dessous (Fig. 23) représente la distribution de la somme des faces des n dés classiques. Graphiquement, nous constatons que plus le nombre de tirages, ou de dés dans notre exemple, augmente, plus la courbe de fréquence se rapproche d'une courbe en cloche symétrique, caractéristique pour la fonction gaussienne (Fig. 24).

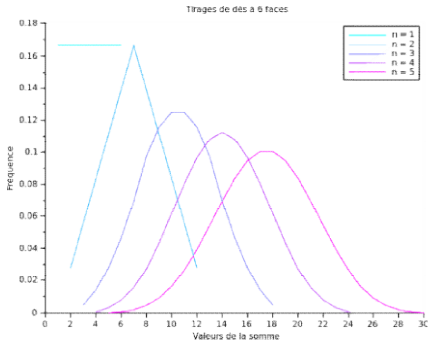


Figure 23 : La somme des faces des n dés classiques © Wikipedia

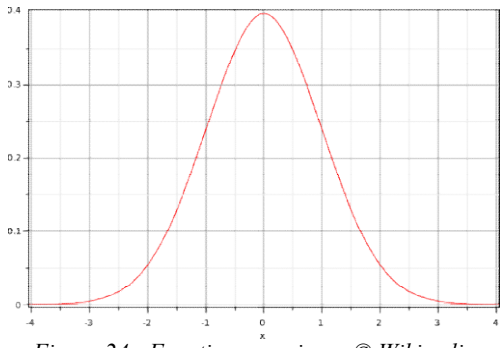


Figure 24 : Fonction gaussienne © Wikipedia

Nous pouvons reformuler ce théorème de la manière suivante : la somme des variables aléatoires indépendantes aura une distribution plus gaussienne que celles des variables d'origine. Comment l'interprétation de ce théorème peut-elle résoudre les problèmes de SAS ?

Nos observations x peuvent être représentées sous forme $x = As$, où A est une matrice de mélange (coefficients scalaires) qui multiplie la matrice des sources s. Connaissant les coefficients de mélange A il suffirait de calculer sa matrice inverse et la multiplier par les observations, $s = A^{-1}x$, mais nous n'avons aucune information sur les coefficients de mélange.

Notons $y = wx$, avec un vecteur w contenant des scalaires quelconques. Si nous développons cette relation, nous avons $y = wx = wAs$. Cela veut dire que y représente des combinaisons linéaires des sources s. Nous pouvons appliquer ici le théorème central limite et dire que la distribution de y sera moins gaussienne quand il sera égal à une des sources s. L'idée est de trouver tels coefficients wA pour lesquels la distribution de y sera la moins gaussienne possible et ainsi, y sera égal à une des sources à quelques indéterminations près. Nous n'avons pas de signaux de source à notre disposition, nous allons donc ne faire varier que les coefficients du vecteur w pour trouver $y = wx$ le moins gaussien possible.

Nous avons vu auparavant qu'il est possible de conclure de manière visuelle sur la gaussianité d'une fonction. Néanmoins nous avons besoin d'une mesure quantitative de cette caractéristique qui pourra être mise en œuvre dans les algorithmes numériques de traitement de données. Dans le livre sur l'analyse en composants indépendantes [2] j'ai trouvé une bonne explication de l'approche générale ainsi que deux critères permettant de mesurer la non gaussianité. La première mesure s'appelle "kurtosis" et pour une variable de moyenne nulle il se calcule comme suit:

$$kurt(y) = E\{y^4\} - 3[E\{y^2\}]^2$$

Pour les distributions gaussiennes le kurtosis est nul, tandis que pour la plupart des variables non-gaussiennes le kurtosis est différent de 0, positif ou négatif. Plus la valeur absolue de kurtosis est supérieure à zéro, moins gaussienne sera la distribution. Kurtosis joue le rôle du critère d'optimisation dans notre approche, cela veut dire que nous cherchons à maximiser ou minimiser sa valeur. Je n'ai pas étudié en détail le deuxième critère - néguentropie - car, comme nous allons voir plus tard, les approches d'ICA ne sont pas vraiment adaptées pour le traitement des données spectrales et la mesure de kurtosis était suffisante pour mes tests et les résultats souhaités.

Une fois le critère trouvé, il suffit de rajouter un algorithme d'optimisation qui cherchera une valeur optimale répondant à nos besoins. Il existe de nombreux différents algorithmes présentés par différents chercheurs. En ce qui me concerne, j'ai découvert et utilisé l'algorithme du gradient conseillé par mon maître de stage.

L'algorithme du gradient s'applique aux fonctions dérivables et consiste à construire une suite de valeurs x_i de manière itérative : $x_{i+1} = x_i \pm \mu . f'(x_i)$ [3]. Le point de départ est fixé au hasard, les signes +- vont maximiser ou minimiser la fonction respectivement. Le pas μ peut être également fixé arbitrairement au début, cependant sa valeur a un impact important sur la convergence de l'algorithme. Les valeurs trop petites vont augmenter le temps de calcul car l'algorithme sera lent, les valeurs trop grandes peuvent provoquer les oscillations et la fonction ne convergera pas. Les valeurs usuelles sont de l'ordre de 0.01, sachant qu'il est toujours possible d'ajuster cette valeur en fonction des résultats obtenus. Pour le critère d'arrêt nous pouvons

choisir un nombre maximal d'itérations ou une valeur très petite pour Δx, cela dépend de nos besoins et connaissances sur la fonction étudiée.

Je vais maintenant détailler un de mes algorithmes appliqué aux signaux sinusoïdaux, exploitant la méthode du gradient et la notion de kurtosis.

Séparation de deux signaux sinusoïdaux

J'ai commencé par créer les signaux sources et les signaux observations (Fig. 25). Les coefficients de mélange A (Eq. 1), qui multiplient les sources pour simuler les observations, sont fixés par des valeurs aléatoires, uniformément réparties entre 0 et 1 à chaque lancement du programme.

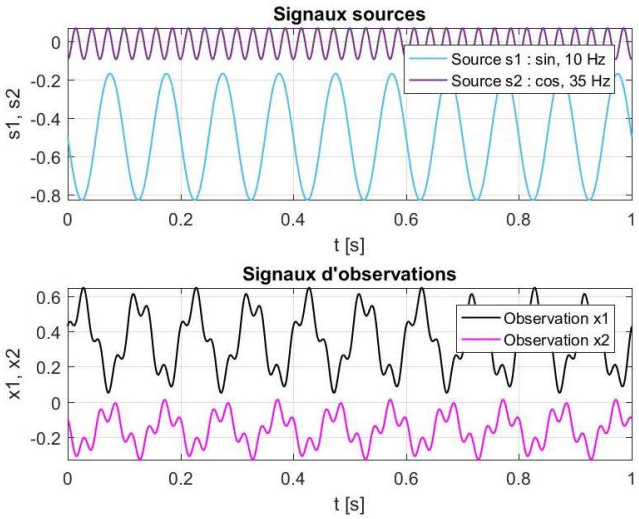


Figure 25 : Signaux et observations créées sous MatLab

$$\begin{bmatrix} x_1(1) & \cdots & x_1(t) \\ x_2(1) & \dots & x_2(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \times \begin{bmatrix} s_1(1) & \cdots & s_1(t) \\ s_2(1) & \dots & s_2(t) \end{bmatrix}$$

Équation 1 : Création des signaux d'observation

Ensuite les signaux observations subissent un prétraitement nécessaire pour la méthode d'ICA :

1. les données sont d'abord centrées en zéro par une soustraction de leurs moyennes (Fig. 26). Cette hypothèse simplifie la théorie de la méthode et les algorithmes ;

```
53    % centering
54 -  x1 = x1 - mean(x1);
55 -  x2 = x2 - mean(x2);
```
Figure 26 : Centrage des données d'observation

2. ensuite les données subissent la procédure de blanchiment (Fig. 27), après laquelle les signaux observations deviennent non-corrélés et chacun a dorénavant l'énergie unitaire. Cette étape permet de limiter les recherches de la matrice de mélange à l'espace des matrices orthogonales[2], ce qui réduit potentiellement le temps de calcul et la complexité des programmes. Les lignes *59-62* correspondent aux calculs de la matrice de blanchiment, qui multiplie par la suite la matrice contenant nos observations centrées (ligne *64*).

```
57 -  M = [x1 x2];          % observations centrées
58
59 -  C = cov(M);           % matrice de covariance de M
60 -  [E, D] = eig(C);      % valeurs et vecteurs propres de C
61
62 -  V = D^(-1/2) * E';    % matrice de blanchiment
63
64 -  M = V*M';             % blanchiment
```
Figure 27 : Blanchiment ou dé-corrélation des données centrées d'observation

[2] Matrices carrées vérifiant ${}^tA\,A = A\,{}^tA = I$, où ${}^tA$ est transposé de A et I est une matrice d'identité.

Après ces deux étapes de prétraitement l'algorithme d'optimisation va chercher une estimation d'une de nos sources. Chacune des lignes de la matrice d'observation M (ligne *64*, Fig. 27) contient les échantillons d'une des observations centrées. Nous allons donc multiplier cette matrice par un vecteur w de taille *2 ligne-1 colonnes*, et faire varier les valeurs de ce vecteur pour rendre le résultat $w^T M$ obtenu le moins gaussien possible[3]. Nous allons procéder de telle manière jusqu'à ce qu'il ne nous reste plus de sources à estimer, et les vecteurs vont former ainsi une matrice de démélange.

En premier lieu j'ai fixé le pas de descente de gradient et le vecteur initial (Fig. 28). Le plus simple est d'initialiser ce vecteur aléatoirement quand nous ne pouvons pas l'estimer.

```
93 -    mu = 0.01;              % "pas" de gradient
94 -    w1 = rand(2,1);         % vecteur initial
```

Figure 28 : Pas de gradient et le vecteur w de départ

Ensuite j'ai mis dans une boucle *while* l'algorithme du gradient (Fig. 29). Pour cette méthode nous avons besoin de la dérivée de notre fonction étudiée. La dérivée de kurtosis $4.sign\left(kurt(w^T M)\right).[E\{M(w^T M)^3\} - 3w||w||^2]$ [2] est stockée dans la variable *grad* (ligne *110,* Fig. 29). Puisque le kurtosis peut prendre des valeurs négatives et positives, il est plus simple de travailler avec sa valeur absolue et de chercher à la maximiser. C'est pour cela que le signe de kurtosis intervient dans l'expression de la dérivée.

```
100 -  while 1
101 -      y = w1' * M;
102 -      y3 = y.^3;
103 -      M1 = M';
104 -
105        % tracer les valeurs de kurtosis à chaque itération
106 -      kurt_old = kurtosis(y);
107 -      plot (iter, kurt_old, '.', 'color', clr1, 'markersize', 15)
108 -      hold on
109 -
110 -      grad = 4 * sign(kurtosis(y)) * ( mean([ M1(:,1).*y3', M1(:,2).*y3']) - (3*w1') );
111 -
112 -      w1 = w1 + (mu*grad');
113 -      w1 = w1 / norm(w1);
114 -
115 -      kurt_new = kurtosis(w1'*M);
116 -      err = abs(kurt_new)-abs(kurt_old);
117 -      if err<1e-6
118 -          break
119 -      end
120 -      iter = iter + 1;
121 -  end
```

Figure 29 : Algorithme du gradient appliqué aux données d'observation

Ensuite le vecteur w doit être mis à jour à chaque itération, comme nous l'avons vu dans la présentation de l'algorithme de gradient. Le signe + (ligne *112,* Fig. 29) signifie que nous voulons trouver un maxima de notre fonction. Dans la méthode de référence choisie [2] ils gardent le vecteur sur le cercle unitaire (sa norme est égale à 1). Cela peut être facilement résolu en divisant le vecteur par sa norme (ligne *113,* Fig. 29).

Enfin, j'ai rajouté une condition d'arrêt (lignes *115-119,* Fig. 29) qui me permet de sortir de la boucle dès que la condition est vraie, sans effectuer de calculs inutiles. Je suppose que l'algorithme a convergé quand la différence entre deux valeurs successives de kurtosis est inférieure à 10^{-6}.

Le vecteur w trouvé ne permet d'estimer qu'une des sources étudiées. Dans le cas où nous avons deux sources à estimer, le deuxième vecteur se calcule très facilement. Rappelons que nous avons restreint notre espace de solutions à l'espace des matrices orthogonales. Un exemple d'une matrice orthogonale est une matrice $\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$. La première colonne étant notre premier vecteur calculé, il suffit de permuter les coefficients et mettre un signe 'moins' devant le premier coefficient pour obtenir le deuxième vecteur, et ainsi l'estimation de la deuxième source.

Ci-dessous (Fig. 30) les estimations des sources de la Fig. 25, obtenues avec mon script MatLab commenté plus haut. Les deux sources sont tracées en violet pointillée, l'estimation 1 en jaune et l'estimation 2 en vert. Toutes les courbes sont tracées sur une même échelle pour pouvoir mieux constater le fait que les estimations sont une image des sources à quelques indéterminations près (facteur d'échelle, signe et permutation d'ordre).

[3] w^T est une transposé de w

Nous pouvons constater que la première estimation trouvée (**e1**) correspond à la seconde source (**s2**), et a une amplitude plus importante que le signal d'origine.

La seconde estimation (**e2**) correspond donc à la source numéro 1 (**s1**) et a également une amplitude plus importante. Cette estimation est aussi un exemple d'une autre indétermination possible - inversion de signe.

En testant le programme plusieurs fois, j'ai pu constater son bon fonctionnement, et plusieurs résultats et cas de figure différents.

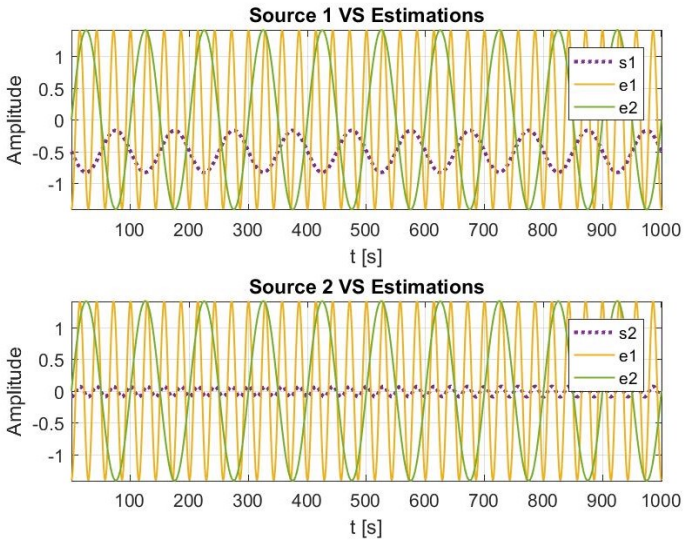


Figure 30 : Comparaison des estimations obtenues et les sources d'origine

Comme mentionné ci-dessus, quand nous n'avons que deux sources à estimer les calculs de la seconde estimation se simplifie. Pour trouver les estimations d'un plus grand nombre de source il serait possible de boucler l'algorithme en changeant le point de départ, une fois le résultat est convergé. Cependant ce n'est pas une solution la plus efficace et élégante. Même avec les points de départ différents, rien ne dit que l'algorithme convergera vers une autre solution et trouvera l'estimation d'une autre source. Pour empêcher les vecteurs de converger vers le même maxima, il est nécessaire de les « orthogonaliser » à chaque itération. Il existe des méthodes différentes pour atteindre ce résultat. J'ai choisi la méthode qui cherche les estimations des sources un par un [2] et l'a testé sur le cas de 3 sources mélangées. Ci-dessous (Fig. 31), dans l'ordre suivant : les sources générées, les observations et les estimations observées. Les lignes de code correspondantes avec la référence de la méthode et les commentaires sont jointes en Annexe (Annexe *Extraits de script MatLab pour ICA, 3 sources à décontaminer*).

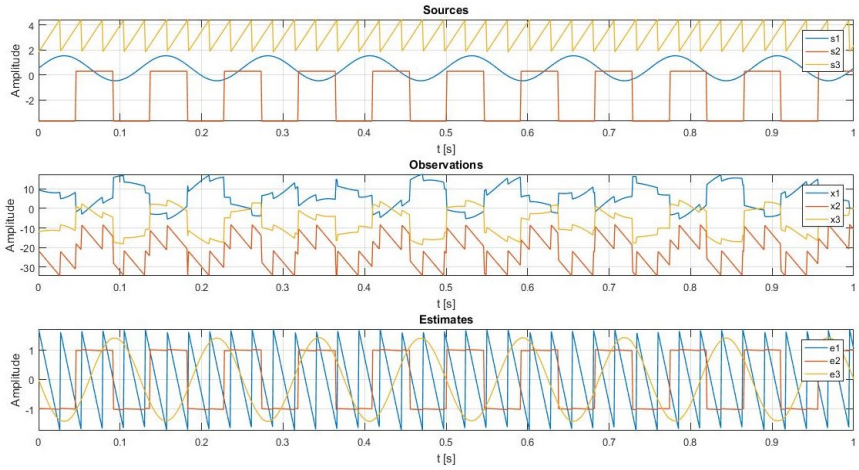


Figure 31 : Conditions de départ et les résultats dans le cas de 3 sources mélangées

Factorisation en matrices non-négatives (NMF)

Les méthodes de Factorisation en Matrices Non-négatives ou NMF (Non-negative Matrix Factorization, en anglais) sont des méthodes basées sur la positivité des données. Par données on entend les observations, les sources et les coefficients de mélange. Un intérêt particulier est accordé à ces méthodes dans le domaine d'astrophysique étant donné la positivité des données acquises par les satellites [1].

Le principe de base est le suivant : étant donné une matrice V représentant les observations, la NMF consiste à trouver deux matrices non-négatives W et H tel que $V \approx WH$.

Les algorithmes NMF sont les algorithmes itératifs qui cherchent généralement à minimiser un critère ou une fonction coût choisie. Le critère classique et le plus utilisé est une distance euclidienne qui se calcule

comme suit : $\frac{1}{2}||V - WH||^2$. C'est la somme de chaque élément de $V - WH$ mis au carré. L'algorithme tourne en boucle tant que cette valeur ne vérifie pas les conditions que nous avons établies.

Différents algorithmes ont été proposés par les scientifiques dans différents champs d'application. Pour les méthodes exploitant la norme euclidienne on distingue trois approches principales suivantes [1].

L'Algorithme du gradient projeté se base sur la même idée, présentée dans la partie sur ICA, mais à la fin de chaque itération il est nécessaire de mettre à jour les matrices. Autrement dit, les projeter sur l'espace des solutions admissibles - tous les éléments négatifs doivent être remplacés par une très petite valeur positive (symbole $[.]_+$ dans Eq. 2) [6]. J étant une fonction coût choisie.

$$W \leftarrow \left[W - \mu \cdot \frac{\partial J}{\partial W}\right]_+$$

$$H \leftarrow \left[H - \mu \cdot \frac{\partial J}{\partial H}\right]_+$$

Équation 2 : Méthode du gradient projeté

L'Algorithme multiplicatif est une extension de l'algorithme de gradient projeté. Le coefficient μ est exprimé en fonction des matrices utilisées (Eq. 3), où $\odot$ et $\oslash$ signifient la multiplication et la division élément par élément respectivement [6].

$$W \leftarrow W \odot ((VH^T) \oslash (WHH^T))$$

$$H \leftarrow H \odot ((W^TV) \oslash (W^TWH))$$

Équation 3 : Méthode multiplicative

Algorithme ALS, Alternating Least Squares en anglais, fixe au début une des matrices, met à jour l'autre, fixe cette seconde matrice et calcule la première, et ainsi de suite. En effet, si nous supposons que nous connaissons une des matrices, W ou H, nous pouvons facilement estimer l'autre, car :

$$V = WH \rightarrow W^TV = W^TWH \rightarrow (W^TW)^{-1}W^TV = H$$

$$V = WH \rightarrow VH^T = WHH^T \rightarrow VH^T(HH^T)^{-1} = W$$

Équation 4 : Approche ALS

Pour cette méthode nous n'avons besoin de fixer qu'une seule matrice de départ, la deuxième étant calculée par la suite. Après chaque mise à jour les matrices doivent être projetées sur l'espace des valeurs positives [6].

Aucune de ces méthodes ne nécessite pas de prétraitement pour les données. La seule condition c'est leur non-négativité.

Comme j'avais déjà mentionné, les méthodes de NMF sont très répandues dans les domaines d'astrophysique et cosmologie étant donné la positivité des données étudiées (spectroscopie, télédétection etc.). De manière générale ces méthodes sont surtout utilisées pour le traitement des images. J'ai traduit et testé ces algorithmes sur un problème de SAS de type 'image' ainsi que sur les données spectrales d'EUCLID dont je parlerai plus loin dans mon rapport. Le premier exemple était un document scanné recto-verso (Fig. 32) où les deux côtés se voient à travers le support, et le but est donc de les séparer pour avoir deux images propres pour le recto et verso.

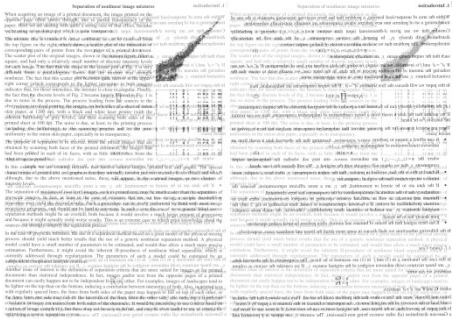


Figure 32 : Document scanné recto-verso avec le support papier semi-transparent

Il est important de remarquer que la deuxième image ne représente pas un vrai scan d'un document. Cette image est une image en miroir d'un vrai document. Cette modification est en effet intentionnelle : si nous prenons une image originale de chacune des côtés de documents, nous allons avoir *quatre* sources mélangées (recto, recto en miroir, verso et verso en miroir). Ce qui transformait notre problème en mélange sous-déterminé (4 sources > 2 observations). Avec l'effet de miroir, nous n'avons que *deux* sources : recto et verso en miroir. Cela ne pose pas de problème pour le traitement et la séparation car à la fin, une fois les estimations trouvées, il suffira d'appliquer l'effet de miroir sur l'image retournée.

Les images chargées sur MatLab sont représentées sous forme matricielle. Ces matrices doivent ensuite être mises sous forme vectorielle (1 ligne ou 1 colonne) pour que l'algorithme puisse les manipuler correctement. Ces vecteurs doivent former une matrice V que nous chercherons à décomposer. Il suffira de réarranger les vecteurs contenants les estimations en matrices pour retrouver la forme des images.

Quant aux matrices de départ, W et H sont remplies des valeurs aléatoires, uniformément réparties entre 0 et 1. Je vais présenter brièvement la mise en œuvre des algorithmes présentés plus haut, ainsi que les résultats obtenus.

Algorithme du gradient projeté

Pour l'algorithme du gradient projeté il a d'abord fallu calculer les dérivées partielles de notre critère $\frac{1}{2}||V - WH||^2$. Les relations obtenues sont implémentées dans le script sur les lignes *35* et *39* (Fig. 33).

Comme mentionné ci-dessus, la seule différence entre l'algorithme du gradient "classique" et le gradient projeté est la projection des résultats sur l'espace des solutions acceptables. Dans notre cas, après chaque calcul de la nouvelle matrice (lignes *36* et *40*, Fig. 33), les valeurs négatives sont remplacées par la valeur ε du MatLab (lignes *37* et *41*, Fig. 33). Cette valeur est égale à $2{,}22.10^{-16}$. Les lignes *43* et *44* (Fig. 33) correspondent aux calculs de la distance euclidienne (critère d'arrêt) entre les observations (les scans recto-verso) et le produit des matrices estimées.

```
32 -  while err > 1e-6
33 -      v = W*H;
34
35 -      gradW = -(V - v) * H';
36 -      W = W - (mu*gradW);
37 -      W = max(W,eps);
38
39 -      gradH = -W' * (V-v);
40 -      H = H - (mu*gradH);
41 -      H = max(H,eps);
42
43 -      e = ((V-W*H).^2)/2;
44 -      err = sum(sum(e));
45
46 -      if iter>200
47 -          break
48 -      end
49
50 -      iter = iter+1;
51 -  end
```

Figure 33 : Implémentation de l'algorithme du gradient projeté

Pour la condition d'arrêt j'ai commencé par une valeur d'erreur de 10^{-6}. J'ai également essayé plusieurs valeurs différentes pour le pas de gradient μ, entre 0,1 et 10^{-6}. En effet, l'algorithme n'arrivait pas à converger vers des petites valeurs d'erreur qui seraient admissibles. Dans un cas la valeur d'erreur oscillait et prenait des valeurs de 10^{10} jusqu'à 10^{60}. Dans l'autre cas l'algorithme convergeait vers un minimum de $\approx 10^8$ en moins de 200 itérations et ne bougeait plus. C'est pour cette raison que j'ai ajouté une condition (lignes *46-48,* Fig. 33) qui terminait l'exécution de la boucle au bout de 200 itérations. Les estimations obtenue sont représentées sur la Fig. 34 (erreur de 7.10^8, 300 itérations).

Cette incapacité de trouver des meilleures solutions avec une erreur proche de zéro peut être dû à la fois aux valeurs du pas de gradient, et aux valeurs des matrices de départ. Depuis les premiers travaux différentes relations mathématiques qui permettent d'optimiser la valeur du pas lors de l'exécution des calculs ont été exposées dans les livres et les publications scientifiques [8].

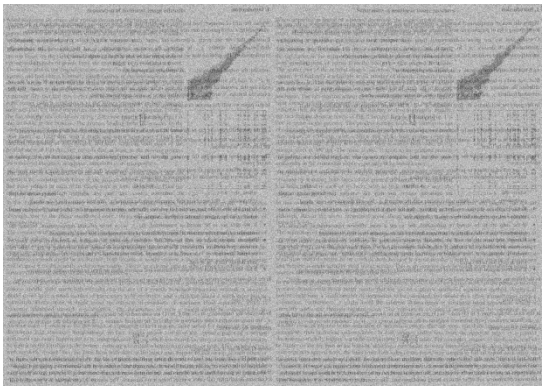


Figure 34 : Estimations obtenues avec la méthode du gradient projeté

Algorithme multiplicatif

L'algorithme multiplicatif a été initialement proposé par D. D. Lee and H. S. Seung [4]. L'idée qu'ils ont développé consiste à fixer une des matrices et essayer de minimiser la fonction coût, tout en respectant la deuxième matrice et ses contraintes (la non-négativité).

Ils ont suggéré les relations de mise à jour (Eq. 3) qui ne peuvent pas augmenter la distance euclidienne, et ainsi faire converger l'algorithme vers une des solutions. Cette méthode présente cependant certains inconvénients. Notamment, il est possible d'obtenir des zéros dans les matrices W et H lors des calculs. La solution la plus simple dans ce cas est d'actualiser les matrices comme nous l'avons vu dans la méthode de gradient dans le but d'éviter les divisions par 0. Plusieurs extensions de l'approche multiplicative ont été également proposées par des chercheurs, incluant d'autres critères d'arrêt et fonctions coûts supplémentaires [1].

Pour mon algorithme (Fig. 35) j'ai ajouté la même règle de mise à jour pour les matrices W et H que dans la méthode du gradient (lignes *33/36,* Fig. 35*)*. Les lignes *32* et *35* (Fig. 35) correspondent aux calculs des matrices selon les formules de Lee et Seung.

```
30 -  while err > 1e-6
31
32 -      W = W.*((V*H')./(W*H*H'));
33 -      W = max(W,eps);
34
35 -      H = H.*((W'*V)./(W'*W*H));
36 -      H = max(H,eps);
37
38 -      e = ((V-W*H).^2)./2;
39 -      err = sum(sum(e));
40
41 -      if iter>200
42 -          break
43 -      end
44
45 -      iter = iter+1;
46
47 -  end
```

Figure 35 : Implémentation de l'algorithme multiplicatif

Les premiers tests de l'algorithme m'ont permis de constater le même problème que dans le cas du gradient : les calculs convergent relativement vite, mais pas vers une valeur souhaitée. L'erreur finale prenait les valeurs de l'ordre 10^5. De même manière j'ai dû forcer la sortie de la boucle au bout de 200 itérations (ligne *41-43,* Fig. 35). Un nombre d'itérations plus grand ne donnait pas meilleures résultats. Par ailleurs, le fait d'avoir le même nombre d'itérations pour les deux algorithmes m'a permis d'effectuer une comparaison équivalente de leurs performances et résultats.

Les premiers résultats obtenus étaient les images blanches. Compte tenu des indéterminations possibles lors de la séparation et les valeurs d'erreurs ($10^5 - 10^6$), je m'attendais à avoir les résultats proches de ceux obtenues avec l'algorithme du gradient. Cependant ce n'était pas le cas.

J'ai décidé donc d'étudier les valeurs des éléments de la matrice H, supposée contenir les échantillons des estimations. Les valeurs étant assez diversifiées, ceci ne me paraissait pas une raison des images blanches observées. Je me suis ensuite intéressée aux fonctions d'affichage et de sauvegarde des images sous MatLab.

La fonction *imwrite,* que j'utilisais pour sauvegarder les estimations en format d'image, prend les valeurs de 0 à 1 et les fait uniformément repartir sur les niveaux de gris de 0 (noir) à 255 (blanc). Les valeurs négatives sont mises à 0, les valeurs supérieures à 1 sont mises à 1. Nos matrices ne contiennent pas de valeurs négatives, ils peuvent cependant avoir les valeurs supérieures à 1. Ce qui était mon cas. Pour résoudre ce problème et pouvoir interpréter les résultats correctement, j'ai décidé de "normaliser" la matrice H. Autrement dit de projeter ses données sur l'échelle de 0 à 1. Voici les lignes de code correspondantes à ma solution suggérée (Fig. 36) :

```
62 -  e1 = (e1-min(e1))./(max(e1)-min(e1));
63 -  e2 = (e2-min(e2))./(max(e2)-min(e2));
```

Figure 36 : Changement d'échelle pour la matrice H contenant les estimations des sources

Après cette étape et réarrangement des vecteurs en matrices, j'ai observé le résultat ci-dessous (Fig. 37 *(a)* et *(b)*). L'erreur dans ce cas était de $2{,}4.10^5$, avec le temps de calcul de 43,8 seconds pour 200 itérations. Nous pouvons constater visuellement que malgré une très grande erreur les images sont interprétables. La première image (Fig. 37 *(a)*) semble être une estimation du côté verso du document. Le fond et le texte du recto sont maintenant séparés du verso en termes de couleurs, ce qui peut réduire le post-traitement supplémentaire à un jeu de couleur entre le fond et le texte. J'ai testé la fonction *imbinarize* de MatLab pour passer l'image en noir et blanc mais la qualité a baissé et l'image est devenu moins lisible (Fig. 37 *(c)*). J'ai ensuite étudié l'histogramme de l'image et essayé d'unifier le texte blanc du recto avec le fond gris. Le résultat était considérablement meilleur (Fig. 37 *(d)*). Il ne reste qu'à appliquer l'effet de miroir qui orientera le texte dans un bon sens.

La deuxième estimation (Fig. 37 *(b)*) correspondait au côté recto du document. L'image n'a pas perdu en qualité et n'avait besoin d'aucun traitement supplémentaire.

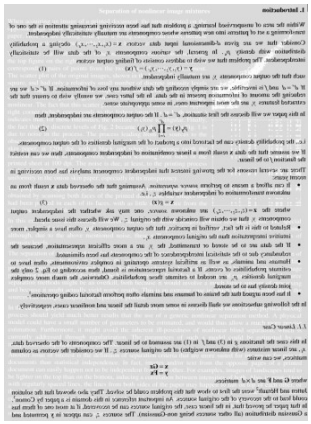
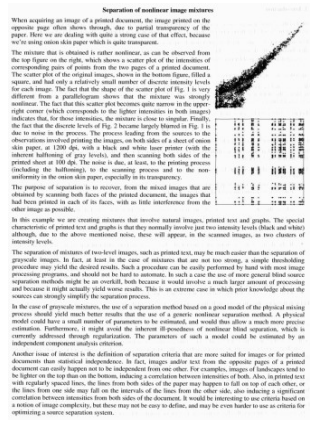
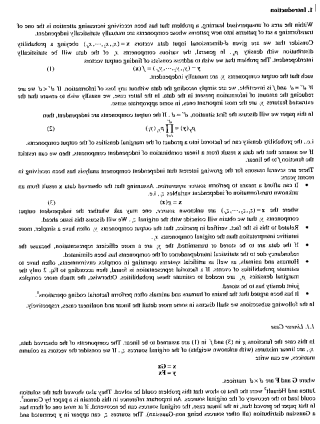


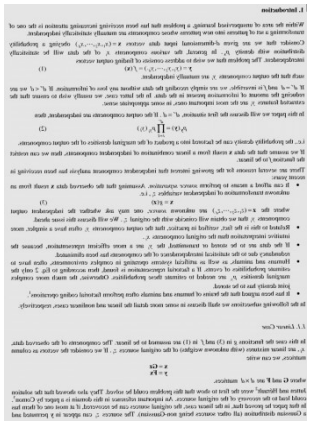
Figure 37 : (a) estimation 1



(b) estimation 2



(c) estimation 1 après binarisation



(d) estimation 1 après l'ajustement des couleurs

Algorithme ALS , Alternating Least Square

L'approche ALS utilise des relations basées sur l'hypothèse qu'une des matrices, W ou H, est connue. Dans mon algorithme (Fig. 38) j'ai décidé de démarrer la boucle avec la matrice W connue (remplie des valeurs aléatoires).

```
27 -  while err > 1e-6
28
29 -      H = inv(W'*W)*W'*V;
30 -      H = max(H,eps);
31
32 -      W = V*H'*inv(H*H');
33 -      W = max(W,eps);
34
35 -      e = ((V-W*H).^2)./2;
36 -      err = sum(sum(e));
37
38 -      if iter>200
39 -          break
40 -      end
41
42 -      iter = iter+1;
43
44 -  end
```

Figure 38 : Implémentation de l'approche ALS

Malgré la positivité de la matrice de départ, les calculs postérieurs peuvent faire apparaître les valeurs négatives. En effet, la matrice inverse d'une matrice positive n'est pas obligatoirement positive. Les calculs itératifs peuvent propager la négativité dans les matrices et fausser ainsi le résultat.

J'ai pu constater cet effet en effectuant un test de l'ALS sans les projections des matrices (lignes *30/33,* Fig. 38). La boucle convergeait au bout d'une seule itération et l'erreur était de l'ordre 10^{-10}. Les images obtenues n'étaient pas pourtant lisibles. J'ai décidé alors de vérifier la positivité des sources estimées - trouver la valeur minimale dans la matrice H . J'ai lancé l'algorithme plusieurs fois, et à chaque nouveau lancement la matrice H comprenait au moins une valeur négative. Ce qui m'a confirmé que la projection des matrices est indispensable dans cette méthode, comme il a été remarqué plus haut.

Comme pour les deux autres algorithmes, j'ai commencé par conditionner la boucle à 200 itérations et mesurer le temps de calcul. Cependant, la valeur d'erreur ne semblait pas de converger au bout de 200 itérations. J'ai donc augmenté le nombre d'itérations pour comparer jusqu'où cette méthode pouvait amener le résultat. Ci-dessous quelques-uns des résultats observés avec les valeurs d'erreurs, nombre d'itérations et temps de calcul (Fig. 39 et 40).

Figure 39 : Estimations avec l'erreur de 660 et le temps de calcul de 48 sec (400 itérations)

Figure 40 : Estimations avec l'erreur de 1105 et le temps de calcul de 25 sec (200 itérations)

En faisant une brève analyse des 3 approches et leurs applications sur les images proposées, nous constatons les différents niveaux de performances des algorithmes :

- l'algorithme du gradient était le moins efficace des trois et n'a pas abouti aux bons résultats. Il serait envisageable de tester les algorithmes d'optimisation de pas de gradient pour pouvoir découvrir plus le potentiel de cette approche;
- l'approche multiplicative a fourni des résultats bien interprétables malgré les valeurs d'erreur plus importantes par rapport aux valeurs souhaitées. Le post-traitement des résultats n'est pas toujours nécessaire, tout dépend des résultats observés;
- la méthode ALS s'est avéré le plus performante des trois en terme du temps de calculs et des valeurs d'erreurs, compte tenu le même nombre d'itérations pour tous les 3 algorithmes. Cependant, comme nous avons pu constater, dans notre exemple une grande valeur d'erreur ne signifiait pas nécessairement un mauvais résultat. De même manière, les petites valeurs d'erreur de l'approche ne fournissaient pas toujours de résultats de *meilleure* qualité (Fig. 39 : l'algorithme n'a réussi à vraiment séparer qu'une des sources à partir les observations).

Si nous comparons la qualité des résultats avec les valeurs de la distance euclidienne correspondantes, ce critère peut paraître en réalité moins qualitatif que nous le voudrions. Pour cette raison les algorithmes plus sérieux et développés possèdent des contraintes supplémentaires en plus de la distance euclidienne ou une autre fonction coût. Cette option permet de réduire le champ des solutions et d'obtenir des estimations plus précises. En outre, les conditions des mises à jour peuvent beaucoup dépendre des informations a priori que nous avons sur les sources et les mélanges.

Après toutes les préparations des algorithmes et les tests j'ai finalement pu me mettre sur les données d'EUCLID. Dans ce qui suit je vais présenter les algorithmes déjà vus auparavant mais appliqués aux spectres des galaxies, ainsi que les résultats de la séparation et une nouvelle analyse des performances des approches vis-à-vis le nouveau type de données.

Données spectrales d'EUCLID

Comme il a été mentionné ci-dessus, le satellite EUCLID est encore en phase de construction et son lancement est prévu pour le Juin 2022. Afin de permettre aux chercheurs d'avancer leurs travaux sur les outils de traitement et séparation des images, les ingénieurs ont développé un logiciel TIPS[4] qui modélise des images représentatives de ce que le télescope Euclid renverra au sol lors de sa mission. Ce sont les images issues de TIPS (Fig. 41) que j'ai utilisé pour la mise en œuvre de mes algorithmes de séparation des spectres.

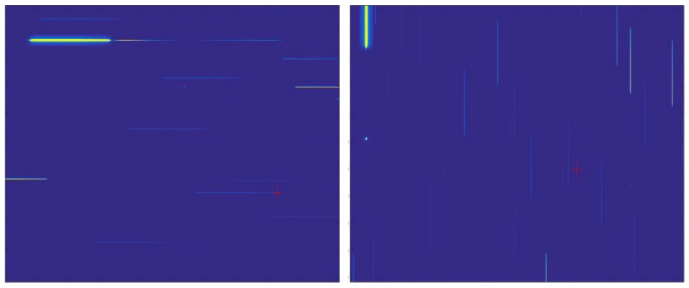


Figure 41 : Exemples d'images issues de TIPS

Je vais présenter ici un des scénarios sur lesquels j'ai pu travailler. Il s'agit d'une image d'une seule direction d'observation avec 2 sources mélangées (Fig. 42 *(a)*). Je parlerai d'abord des principes généraux de travail avec telles données spectrales, ensuite je résumerai les algorithmes appliqués à ces données, et enfin je vais analyser les performances des méthodes et la qualité des résultats.

Comme nous l'avons vu dans la partie sur les instruments d'EUCLID, les grismes étalent la lumière des corps célestes formant ainsi leurs spectres. Notons que le spectre d'un objet occupe plusieurs lignes de pixels (Fig. 42 *(b)*). Cependant, si nous avons un seul objet, son spectre est le même sur chacune des lignes à un facteur

[4] Pour "This Is a Pixel Simulator" en anglais

d'échelle près : il sera plus brillant sur la ligne centrale et moins brillant sur les autres lignes. Chaque ligne peut être donc considérée comme *une* observation.

Si maintenant nous étudions deux corps dont les spectres sont superposés. Sur l'image zoomé nous pouvons distinguer deux "verticales" brillants (Fig. 43, entourés en rose) correspondants à deux corps célestes. Si nous prenons la ligne n°6 avec la raie brillante du corps situé à gauche, sur cette ligne à droite il y a une contribution du spectre de l'autre objet entourée en rectangle jaune. De même manière, la ligne n°7 comprenant « le centre » du deuxième objet est contaminée par une partie du spectre du premier objet (rectangle jaune).

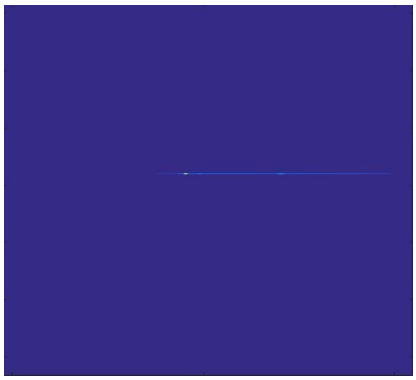
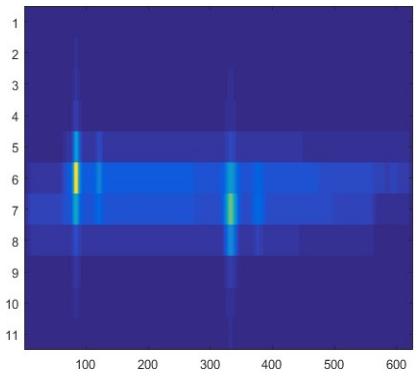


Figure 42 : (a) spectres de deux corps célestes observés, vue d'ensemble



(b) vue approché de deux corps étudiés

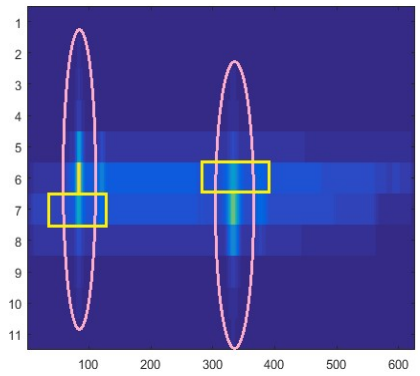


Figure 43 : Détermination des lignes centrales et contaminations

De préférence nous choisissons les lignes avec les raies les plus brillantes pour avoir le maximum d'information et de précision. Quant aux largeurs des intervalles, pour les données réelles d'EUCLID les chercheurs ont des chiffres exacts sur les pixels de début et de fin des spectres. Ceci va leur permettre de traiter les données correctement avec une meilleure précision.

Pour mes algorithme j'ai choisi les lignes n°6 et 7 en tant que les observations. Les lignes de code correspondant aux méthodes de séparation de sources n'ont pas changé. Pour préparer les données, j'ai récupéré la position du pixel le plus brillant, et à l'aide de ces coordonnées j'ai déterminé visuellement les autres lignes à traiter.

Sur la figure ci-dessus (Fig. 44) les spectres mélangés obtenus de deux lignes d'observations choisies :

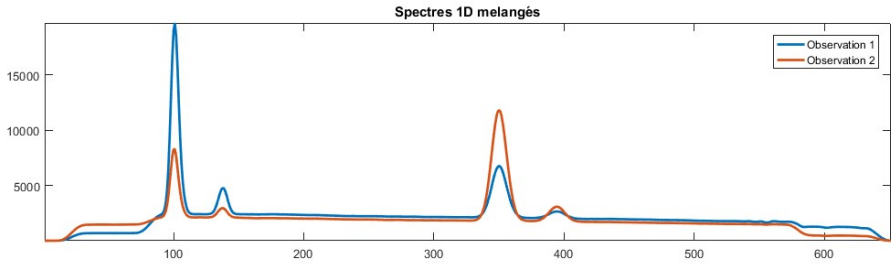


Figure 44 : Spectres de deux corps étudiés en fonction de la position du pixel

Ci-dessous les résultats obtenus avec les différents algorithmes ainsi que leurs performances. Pour ce scénario j'ai également eu les spectres sources de deux objets étudiés. J'ai donc pu comparer les estimations issues de mes algorithmes avec les vrais spectres.

ICA + l'algorithme du gradient projeté

Le temps moyenne de calcul pour cette méthode était de $1{,}8\,sec$. Les résultats (Fig. 45) étaient exploitables, avec quelques petites imperfections ce qui est parfois inévitable. Nous constatons que les deux piques de départs (Fig. 44) sont maintenant séparés et chacun d'eux ne contient pas de contamination de l'autre. Les données sources sont tracées en rouge, les estimations en noire ; la première source et son estimation sont sur le graphe du haut, la deuxième source avec son estimation sur le graphe du bas.

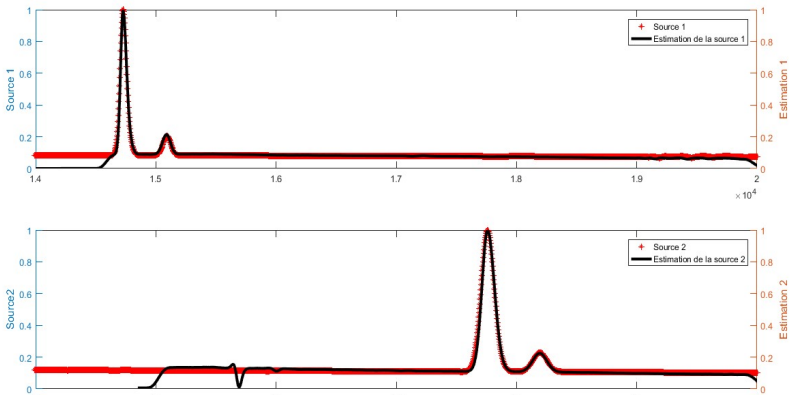


Figure 45 : Estimations issues de l'ICA (gradient projeté)

NMF + l'algorithme du gradient projeté

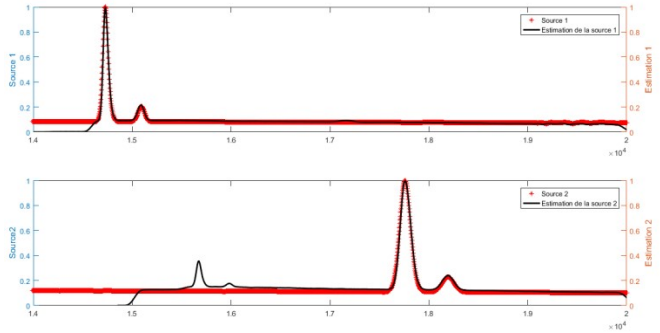


Figure 46 : Estimations issues du NMF (gradient projeté)

Valeurs moyennes, 25 essaies	
Temps de calcul	1,33 sec
Erreur	8.10^{-4}
Nombre d'itérations	967 (condition d'arrêt : erreur<10^{-6}, sinon 1000 itérations)

Certains résultats n'étaient pas si « propres » que souhaité (exemple Fig. 46). Pourtant, la majorité des résultats de cet algorithme étaient interprétables. Dans plus de la moitié des essais effectués l'algorithme convergeait jusqu'à la valeur d'erreur souhaitée.

NMF + algorithme multiplicatif

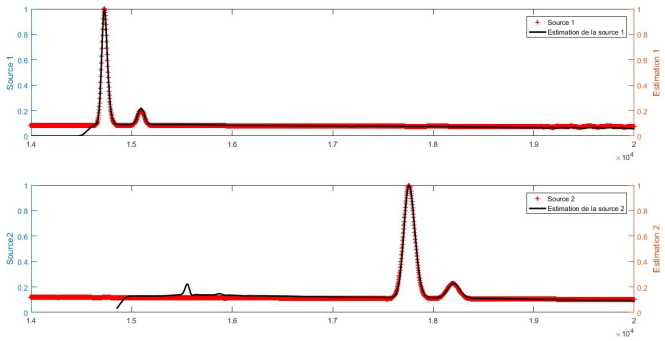


Figure 47 : Estimations issues du NMF (méthode multiplicative)

Valeurs moyennes, 120 essaies	
Temps de calcul	1,42 sec
Erreur	1541
Nombre d'itérations	926 (condition d'arrêt : erreur<10^{-6}, sinon 1000 itérations)

L'algorithme a convergé vers une valeur souhaitée en moins de 1000 itérations dans 17 de 120 essais. Les valeurs d'erreur obtenues au bout de 1000 itérations sont assez diversifiées. Les résultats étaient interprétables même avec les valeurs d'erreur supérieures aux valeurs souhaitées. Les imperfections des estimations sont moins importantes que celles des résultats précédents (le cas du gradient projeté).

NMF + algorithme ALS

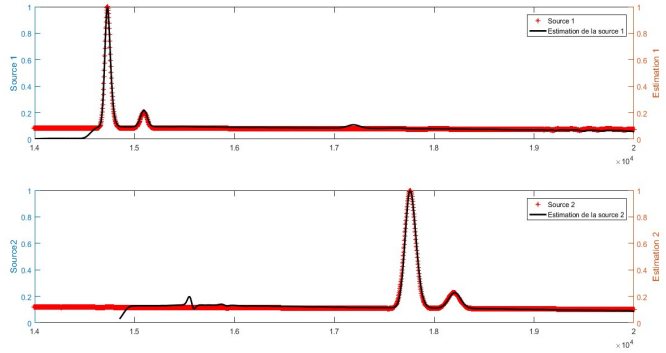


Figure 48 : Estimations issues du NMF (approche ALS)

Valeurs moyennes, 37 essaies	
Temps de calcul	2,68 *sec*
Erreur	33,6
Nombre d'itérations	1974 (tests avec 1000, 2000 et 3000 itérations tout inclus)

Malgré la vitesse et les valeurs d'erreur plus petites que celles issues de deux autres méthodes, les résultats ne correspondaient pas toujours aux estimations attendues. Au moins dans un tiers des cas les données étaient peu crédibles (une estimation correspondant à une des observations, la deuxième estimation étant .presque nulle, par exemple). Les imperfections étaient également moins importantes que celles des résultats issus de la méthode du gradient projeté.

En observant tous les résultats, la qualité et les performances, j'ai jugé les méthodes appliquées assez efficaces et capables de fournir des solutions satisfaisantes dans les cas des problèmes assez simples. C'est-à-dire dans les cas des données non bruitées, avec un nombre de sources à estimer pas trop grand. Une continuation envisageable de ces tests serait d'appliquer ces algorithmes sur les données bruitées, par exemple, ou sur les observations avec plus de deux sources.

En ce qui concerne les méthodes d'ICA, la plupart du temps elles ne peuvent pas être appliquées sur les données spectrales à cause de la corrélation possible entre les spectres des corps célestes. En effet, deux galaxies ou étoiles ayant approximativement le même âge et contenant les mêmes éléments chimiques vont avoir les mêmes spectres. Du point de vue statistique ces spectres seront considérés comme les données corrélées et, donc, non indépendantes. C'est pour cette raison que les chercheurs se focalisent plus sur la méthode NMF dont l'hypothèse de la positivité des données est toujours vérifiée dans le cas de la spectroscopie. À cela s'ajoute que les données issues d'ICA auront éventuellement besoin d'un post-traitement pour rendre les données positives et les exploiter correctement, ce qui n'est pas toujours nécessaire dans le cas des approches NMF car les données restent positives tout au long des calculs.

Pour les méthodes de NMF il serait intéressant d'étudier les extensions possibles des approches appliquées : fonction coût/critère d'arrêt supplémentaire, optimisation du pas de gradient etc., pour observer la qualité des résultats. Nous ne pouvons pas également conclure sur la robustesse de ces méthodes contre le bruit. Il serait intéressant d'essayer de décontaminer les données bruitées, qui représentent aussi une situation réelle qui peut arriver lors du sondage du ciel par les instruments d'EUCLID. Le bruit dans ce cas peut être la lumière des étoiles plus proches de la caméra et donc plus brillante, ou les bruits de nature électronique (issus des capteurs).

Par ailleurs, dans le cas des algorithmes en leur version simple il peut être nécessaire d'ajouter quelques opérations de post-traitement afin de mettre les solutions en forme. Pour certaines algorithmes les contraintes qui y sont implémentées en plus des critères principales peuvent éliminer les étapes de mise en forme, ainsi que rendre les solutions plus propres et précises. Ce point peut être également vu comme un critère de qualité de tel ou tel méthode ou programme : l'algorithme de séparation des sources est-il suffisant et fournit-il les résultats exploitables directement, ou aurons-nous besoin d'ajouter des opérations intermédiaires entre les étapes de démélange et l'utilisation des données.

Pour conclure, les programmes que j'ai écrits ont accompli le « travail » demandé et répondu à notre premier besoin – séparer les sources à partir des mélanges proposées. En plus des applications présentées dans le rapport, j'ai pu tester la séparation des mélanges de plusieurs signaux audio et des signaux biomédicaux et obtenir des résultats satisfaisants. Ainsi, les algorithmes mis en œuvre s'est avéré assez universel et fournissant des résultats pertinents.

Bilan et analyse de la réalisation des objectifs

Réalisation des objectifs définis

Si je reviens sur les étapes fixées au début du stage qui présentaient un plan structuré de mon travail, je peux dire qu'en principe je l'ai suivi du début jusqu'à la fin :

- ☑ j'ai étudié un bon nombre de sources différentes sur les sujets de mon stage (séparation des sources, spectroscopie etc.) et appliqué ces nouvelles connaissances dans la suite de mon travail. J'ai cité dans la bibliographie les références principales que je consultais fréquemment durant le stage ;
- ☑ non seulement j'ai acquis des nouvelles notions du domaine de traitement du signal et de l'image qui peuvent m'être utile pour mes études et projets en Master (algorithmes d'optimisation, indépendance des données, méthodes de séparation des sources) mais j'ai pu aussi approfondir mes connaissances et compétences obtenues en L3 (codage sous MatLab, opérations de base sur les images et les série de données, notions de base de statistique et probabilité) ;
- ☑ j'ai implémenté les méthodes de SAS étudiées sous MatLab pour plusieurs séries de données différentes (signaux audio, signaux sinusoïdaux, images, données spectrales) et obtenu des résultats satisfaisants qui m'ont permis de constater le bon fonctionnement de mes programmes ;
- ☑ j'ai pu découvrir le domaine de spectroscopie spatiale, travailler sur les données spectrales (même si ces dernières n'étaient pas que des simulations) et apprendre plus sur les problématiques de traitement de données spatiales ;
- ☑ enfin, j'ai essayé d'être autocritique et analyser mon travail effectué et ses résultats sous un autre angle. Cela m'a permis de m'évaluer et réfléchir aux améliorations que je pourrais apporter à mon travail et à la suite que je peux donner à mes recherches d'informations et compétences acquises.

Bilan humain et organisation du travail

Lors de mon stage j'ai intégré l'équipe SISU qui comprend une vingtaine de personnes « permanentes » (chercheurs, enseignant-chercheurs, doctorants et post-doctorants) ainsi que les stagiaires de Master et Licence. En soi je ne travaillais pas *avec* quelqu'un, c'est-à-dire mon travail ne dépendait pas des résultats des autres et le travail des autres n'affectait pas mon travail. Cependant, je communiquais avec les autres stagiaires avec qui je partageais l'espace de travail, ainsi qu'avec les autres membres du groupe SISU pendant des poses ou des événements organisés par l'IRAP. Cette intégration m'a permis de faire connaissance avec les gens du domaine de traitement du signal et de l'image qui ont partagé avec moi beaucoup d'informations sur le travail dans un laboratoire, la vie des doctorants et les études dans le domaine. Ces échanges m'ont également permis d'apprendre plus sur le domaine dans lequel j'aimerais travailler par la suite et d'enrichir ma culture générale.

Durant mon stage M Hosseini m'aidait dans la progression des mes tâches. Nous organisions des rendez-vous où je pouvais exposer mes résultats et tenir M. Hosseini au courant de l'avancement de travail. Si j'avais une question, il était toujours disponible pour me répondre ou m'aider à comprendre tel ou tel concept de la problématique de mon sujet de stage. Je me renseignais également auprès des autres stagiaires sur les questions de MatLab ou leurs expériences des études en Master. Les échanges avec M. Hosseini m'ont également permis de découvrir le contexte du travail dans le spatial et dans le domaine de recherche.

Bilan personnel

Le contexte du travail individuel durant le stage m'a fait réfléchir sur mes méthodes de travail, mes habitudes et mon organisation.

En prenant du recul je ne trouve pas que j'ai réussi à répartir équitablement mon temps disponible sur les différentes tâches et étapes de travail confié. La complexité de certaines notions et le fait de découvrir tout un nouveau champ de connaissances et applications qui m'intéressaient de plus en plus me faisait passer plus du temps sur les programmes et données assez simples qui ne devaient être qu'une partie préliminaire pour le reste du stage. Ces parties m'étaient sans doute indispensables pour la meilleure compréhension du sujet.

Cependant j'aurais pu leur consacrer le minimum du temps nécessaire et m'avancer plus sur les données spectrales tout en découvrant le sujet de stage et en approfondissant les connaissances acquises de la même manière qu'avec les applications plus simples. Cette expérience m'a motivé à chercher à améliorer mes méthodes et organisation de travail, ainsi que de travailler sur ma capacité de gestion de temps et de priorités.

Conclusion

Malgré les difficultés de compréhension du contexte de stage que j'ai rencontré au début et un manque de temps pour le travail sur les données spectrales, je suis contente des résultats obtenus et du travail effectué. Le stage était une grande opportunité pour moi de rencontrer et faire connaissance avec les gens du domaine dans lequel j'aimerais travailler et d'améliorer mes compétences en communication. Ce stage a également confirmé mon choix du parcours pour la suite de mes études et ma carrière.

C'était un grand investissement dans mes connaissances et expérience que je pense réaliser par la suite dans ma vie professionnelle.

Annexes

L'effet Doppler

L'effet Doppler dit que la fréquence d'onde change si la distance entre l'émetteur et le récepteur varie au cours du temps (Fig. 49).

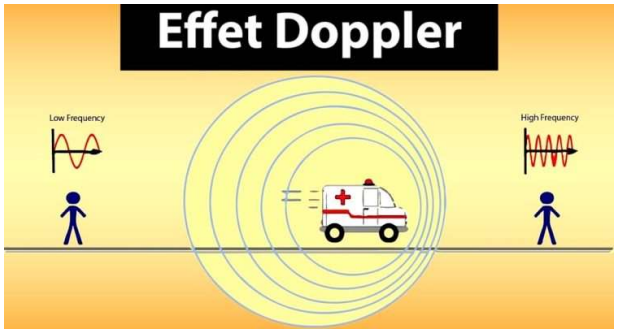


Figure 49 : Illustration de l'effet Doppler© trustmyscience.com

Si la source s'éloigne, les ondes s'étirent de plus en plus en augmentant leur longueur. S'il s'agit d'un spectre électromagnétique la raie va être décalée à droite vers le rouge, dans le sens croissant de longueur d'onde *(redshift)*. S'il s'agit des ondes sonores, un véhicule qui s'éloigne de nous, nous entendons un son plus bas et grave.

En revanche, si la source s'approche, les ondes se rétrécissent en diminuant leur longueur. Dans ce cas la raie va être décalée maintenant vers le bleu, dans le sens décroissant de longueur d'onde *(blueshift)*. Dans l'exemple de véhicule, si celui-là s'approche de nous, nous entendons le son plus aigu.

Plusieurs corps célestes peuvent s'éloigner de nous ou s'approcher. En 1929 les astronomes américains Edwin Hubble et Milton Humason énoncent la Loi de Hubble, qui dit que les galaxies s'éloignent les unes des autres à une vitesse approximativement proportionnelle à leur distance.

Extraits de script MatLab pour ICA, 3 sources à décontaminer

Auteur de script - Daria MALIK, méthode trouvée dans [2] , pages 194-196

```
92 -    moy = mean(X);                    % centering
93 -    moy = repmat(moy, sizeX(1), 1);
94 -    X = X - moy;
95
96 -    C = cov(X);
97 -    [E, D] = eig(C);
98 -    V = D^(-1/2) * E';
99 -    X = V*X';                         % blanchiment
```

```
133 -    w = [];                          % future matrice de demélange
134 -    mu = 0.01;                       % "pas" de gradient
135 -    iter = 1;
136
137 -   for p=1:m                         % m = nombre de sources à estimer
138 -       wp = rand (m,1);              % vecteur initial
139 -       if p==1                       % recherche d'un premier vecteur de demélange
140 -           for iter=1:100
141 -               y = wp' * X;
142 -               y3 = y.^3;
143 -               X1 = X';
144 -               grad = 4 * sign(kurtosis(y)) * ( mean([ X1(:,1).*y3', X1(:,2).*y3', X1(:,3).*y3']) - (3*wp') );
145 -               wp = wp + (mu*grad');
146 -               wp = wp / norm(wp);
147 -           end
148 -       end
149 -       if p>1
150 -           for iter = 1:100
151 -               y = wp' * X;
152 -               y3 = y.^3;
153 -               grad = 4 * sign(kurtosis(y)) * ( mean([ X1(:,1).*y3', X1(:,2).*y3', X1(:,3).*y3']) - (3*wp') );
154 -               wp = wp + (mu*grad');
155
156 -               for j=1:p-1                          % orthogonalisaiton
157 -                   wp = wp - ((wp'*w(:,j))*w(:,j));
158 -               end
159
160 -               wp = wp /  norm(wp);
161 -           end
162 -       end
163 -       w = [w wp];  % les vecteur trouvés sont stockés dans une matrice
164 -   end
```

Bibliographie

[1] Inès MEGANEM, *Méthodes de Séparation Aveugle de Sources pour l'imagerie hyperspectrale. Application à la télédétection urbaine et à l'astrophysique*, Traitement du signal et des images, Toulouse, Université Toulouse III - Paul Sabatier, 2012, pages 7-8

[2] Aapo Hyvärinen, Juha Karhunen et Erkki Oja, *Independent Component Analysis*, John Wiley & Sons, 2001, pages 160-196

[3] Christopher Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995, chapitre 1, pages 17-28

[4] Daniel D. Lee et H. Sebastian Seung, *Learning the parts of objects by nonnegative matrix factorization*, *Nature*, 401, 1999, pages 788–791 (1999)

[5] Ahmed SELLOUM, Shahram HOSSEINI, Yannick DEVILLE et Thierry CONTINI, *Mixing model in slitless spectroscopy and resulting blind methods for separating galaxy spectra,* Italy, IEEE International workshop on machine learning for signal processing, 13-16 sept. 2016

[6] Ngoc-Diep Ho, *Nonnegative Matrix Factorization Algorithms and Applications*, Ingénierie Mathématique, Louvain, Université Catholique De Louvain École Polytechnique De Louvain, 2008, pages 49-63

[7] Kenneth D. Lawrence, Stephan Kudyba et Ronald K. Klimberg, *Data mining methods and applications*, Auerbach Publications, 2014, pages 87-95

[8] Andersen Ang, Non-*negative Matrix Factorization via (normal) Projected Gradient Descent* (pdf)

[9] CANDELS Spectroscopy: The Infrared Grism, http://candels-collaboration.blogspot.com/2012/07/candels-spectroscopy-infrared-grism.html

[10] EUCLID, https://euclid.cnes.fr/fr

[11] IRAP Astrophysique et Planétologie, http://www.irap.omp.eu/

[12] CNRS, http://www.cnrs.fr/