Méthodes pour l'analyse des champs profonds extragalactiques MUSE : démélange et fusion de données hyperspectrales ; détection de sources étendues par inférence à grande échelle.

Raphael Bacher

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Signal, Image, Parole, Télécoms (SIPT)**

Arrêté ministériel : 25 mai 2016

Présentée par
Raphael BACHER

Thèse dirigée par **Olivier MICHEL**,
codirigée par **Florent CHATELAIN** et **Roland BACON**

préparée au sein du **Laboratoire Grenoble Images Parole Signal Automatique**,
dans **l'école doctorale d'électronique, électrotechnique, automatique et traitement du signal (EEATS)**

Méthodes pour l'analyse des champs profonds extragalactiques MUSE : démélange et fusion de données hyperspectrales ; détection de sources étendues par inférence à grande échelle.

Thèse soutenue publiquement le **8 novembre 2017**,
devant le jury composé de :

Monsieur Olivier MICHEL
Professeur, Grenoble INP, Directeur de thèse
Monsieur Hervé CARFANTAN
Maître de Conférences, Université Toulouse 3, Rapporteur
Monsieur André FERRARI
Professeur, Université Nice Sophia Antipolis, Rapporteur
Monsieur Christophe COLLET
Professeur, Université Strasbourg, Président
Monsieur Roland BACON
Directeur de Recherche, CNRS Délégation Rhône Auvergne, Co-directeur de thèse
Monsieur Florent CHATELAIN
Maître de Conférences, Grenoble INP, Co-directeur de thèse
Madame Céline MEILLIER
Maître de conférence, Université de Strasbourg, Invitée

Remerciements

Bonjour ami lecteur, arrivant par hasard ou à dessein sur ces pages.

Finir une thèse c'est notamment l'occasion de prendre le temps de se retourner et de regarder tout le chemin parcouru. Voici donc comme il est de raison quelques mots décrivant ma profonde affection pour tous ceux et celles qui de ce périple furent les valeureux compagnons.

Commençons bien sûr par ceux qui ont rendu cette thèse possible.

Je puis en effet avec fierté claironner que pendant cette thèse, la FORCE [1], la Florent, Olivier, Roland, Compagnie d'Experts, fut toujours avec moi. Un très grand merci à tous les trois, pour tout ce que vous m'avez apporté. Vous avez formé une équipe très complémentaire, à la fois compétente, à l'écoute et bienveillante. La réussite de cette thèse vous doit beaucoup.

Merci également à l'ensemble des membres du jury, Christophe, André, Céline et Hervé pour avoir évaluer mon travail, j'ai beaucoup appris de vos remarques et conseils.

Il est à présent temps de remercier tous ces compagnons de route qui ont permis que cette thèse se passe dans un cadre particulièrement agréable.

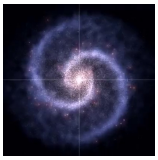
Comment ne pas commencer par Céline, présente au début et à l'arrivée. J'ai débuté en décryptant ton code, tu as fini en évaluant mes méthodes. Je suis très heureux et fier d'avoir été le premier à t'avoir compté dans son jury ! Merci pour m'avoir si bien accueilli dans ce fameux groupe SICOM, moi pauvre Centralien égaré.
De Céline à Quentin, il n'y a qu'un pas, que je franchis allégrement. Quentin, tu resteras le grand blond (à la chaussure noir ?), venu me chercher dans ma salle stagiaire le premier jour : "Salut on va manger, si tu veux te joindre à nous". Life was never the same after. Merci pour tout et le reste.

Et puis merci bien sûr à tous les anciens :

Aux deux Timothées, aussi différents mais aussi précieux l'un que l'autre, tels les deux Ajax. Timothée avec deux yeux et une bouche, avec au moins autant d'idées de projets à la seconde que moi, mais qui en plus les réalise ! Merci pour le poisson. Timothée 2G, à l'enthousiasme débordant, et au coffre à costumes bien rempli, merci pour les innombrables discussions sur les drones, le cinéma, etc... J'avoue, je l'espère, que finalement Djal va te faire craquer et revenir définitivement sur Grenoble...

Aux gardiens en chef du seven wonders, et figures tutélaires du groupe : Aude, Robin et désormais Clotilde. Aude, merci pour ton enthousiasme et ton talent théâtral mais également pour avoir fini par lâcher les sous du Gipsadoc (le compte en banque est resté à ton nom jusqu'en

1. oui, trouver des acronymes fait partie intégrante du processus de recherche : faire de la science n'est-ce pas en effet chercher à nommer le réel ?



2015 quand même...). Merci Robin pour ces parties de tennis disputées, et toutes ces discussions geek tout aussi endiablées.
A la seconde petite famille, Arnaud, Hélène et Nestor. Merci pour tout ces bons moments, en Islande notamment.
A Pascal, longtemps seul camarade de combat pour défendre certaines oeuvres du septième art. Mais je crois qu'on a fini par imposer la ouiche lorraine.
A Manu, pour être Manu.
A Cindy, au rire aussi dévastateur que communicatif.
A Taia, l'enfant des îles, pendant longtemps un simple nom sur un rapport de stage, avant une rencontre ... étonnement au bar.
A Lucas, l'homme aux doigts de fée lorsqu'ils rencontrent des cannes de babyfoot. Te voir jouer et plonger pour récupérer les balles aux périls de tes phalanges c'était beau. Beau et terrible à la fois.
A Victor, le Toulonnais, et Paolo, il rossonero, (e tutta la famiglia) pour avoir choisi des équipes qui permettent de les chambrer régulièrement.

Merci également aux camarades de cordée :
Marc, l'homme le plus endurant du monde. C'en est presque énervant.
Pierre, le bricoleur de génie.
Alexandre M., successeur avec brio à la tête du Gipsadoc, arrivant à gérer tous les pots de Noël depuis Toulouse.
Miguel, enfin, le plus français des espagnols, ou le plus français des espagnols je sais plus. Merci pour ces années de co-bureau et de co-amitiés et courage pour ces derniers mois !

Sans oublier tous les autres, Guillaume et Carole pour m'avoir accueilli dans le bureau "au fond derrière la photocopieuse" (hein Guillaume, quelle histoire ça aussi...), Benoit et Benoit, Sylvia, Alexis, Emmanuelle, Romain, Mai, Saloua, Francesca, Jérémie, ...

Et puis merci à la deuxième génération : Marielle (et Albin), Camille, Marion (et Fabrice), Théo, Florent... Merci d'avoir assuré la relève à la coinche, d'avoir organisé moult parties de volley... Et Marielle et Marion, on s'est bien régalé à Kos !

Merci enfin à ceux de la dernière heure, qui ont apporté un vent de fraîcheur, et ont permis que cette fin de thèse se déroule autant que possible dans la décontraction et la bonne humeur !
Jeanne. Déjà pendant ton stage tu as fait roulé un train de fraîcheur sur les rails de la routine. Et maintenant que tu es capable de comprendre cette référence, tu as encore franchi une étape vers la classe. Mais n'oublie pas, un grand pouvoir implique de grande responsabilité. Fais-en bon usage.
Aziliz, la joueuse de foot farouchement bretonne, dont le rire a rappelé aux murs du Gipsa les plus grandes performances de Cindy.
Et également, Pedro, l'homme de Rio, Kevin, la classe inhérente à un Watson londonien, Thibaut le citoyen éclairé, ...
Sans oublier dans cet incessant inventaire, Pierre, mon très valeureux stagiaire, qui a su travailler sans relâche, pendant que je vaquais à diverses tâches, en Islande, à Porquerolles ou ailleurs...

But also, the new internationals, Maria y Edurne, Geoff, Ondrej, ... Thanks guys for these few months !

A vous, encore jeunes padawans, je n'aurai que ce conseil de nouveau vieux con : courage le chemin est long mais la voie est libre...
Et à tous, pour les bons moments, les coinches, les soirées jeux, les soirées crêpes, merci.

Je voudrais également saluer Gipsadoc et toutes ses équipes pour tous leurs efforts : Maël et Maëlle, Céline, Pascal, puis Marc, Pierre, Miguel, Alex M., et Alex H. (également pilier du foot@gipsa), Sophie (digne héritière d'une présidence centralienne !) et enfin Thibaut, Rémi, Nicolas, Raul... Ce fut un plaisir de participer à ces évènements avec vous. Bonne continuation aux nouvelles équipes !

Merci également à toute l'équipe CICS, incontestablement du labo la meilleure équipe, j'ai beaucoup appris de ces moult réunions.

Enfin merci à tous les services, Lucia, Martine, Marielle, Akila,... pour nous permettre de travailler dans d'aussi bonnes conditions !

Cette thèse m'a également donné la belle opportunité de travailler dans un deuxième laboratoire, à Lyon. Merci donc à la bande de joyeux astronomes du CRAL :
Benjamin l'organisateur en chef de tout événement, Peter, Vera, et tous les autres (notamment pour ce week-end astro-montagne-fossiles inoubliable !).
Mais également à Mylène et Sylvie pour avoir géré tout du long notre situation particulière, beaucoup au loin, un peu à Lyon. Merci Laure et Simon pour tout le travail en amont et en aval sur le code. Courage Laure, les Verts ne peuvent que remonter ;). Thanks also Mohammad for your help and the discussions on the deblending problem.
Merci enfin Jean-Baptiste, Carole et Floriane, compagnons au plus près de cette aventure, ainsi que leurs encadrants, pour tous ces échanges enrichissants. C'était une chance immense de partager cette expérience ensemble !

Enfin avant de conclure, parlons désormais des nombreux amis qui ont transformé ces trois ans de thèse en trois de vie et de projets.

Merci donc aux théâtreux de la première heure.
Manu, Quentin, Aude et Katia, puis Tim, Chloé, Tim, Pascal, Nelson et Marine, qui ont su souffrir mes excentricités de metteur en scène.
Mais également à tous les anciens U1, Philippe, Gabriel, Alizée, Colin, Thibaut, Marie...
Nelson, enfin, colocataire fidèle, merci pour tous tes éclairs de génie et incroyables projets, mais aussi pour partager avec moi les looses du dimanche soir...
Et Antoine, bien sûr, camarade de Khôlle depuis déjà 10 ans ! Allez c'est aussi bientôt ton tour, tu verras ça change rien mais c'est cool quand même ;) !

Et ces trois années furent aussi l'occasion d'observer les étoiles sans pression... Merci donc à Marie, Philippe, Nathan, Baptiste, Manuela, Guillaume, Antoine, Alex, Marc, et Tiphaine. Le plus difficile en thèse c'est d'arriver à lâcher prise alors merci pour ces semaines hors du temps,

sous les cieux étoilés, à profiter du moment.

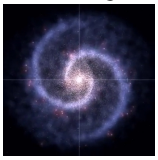
Il est enfin le temps, de conclure, par ce qu'il y a de plus sûr. A toute ma famille, mes parents, ma soeur, merci pour tout, et notamment de me soutenir sans réserve dans tous mes projets.

Voilà donc ami lecteur, le bel aéropage, qui a permis l'écriture de ces pages. Je te souhaite donc, pour enfin conclure, une agréable lecture.

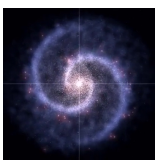
Live long and prosper. \\//

Table des matières

Table des figures



Liste des tableaux



Liste des Algorithmes

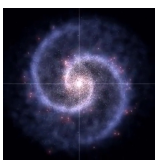
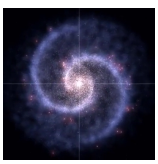


Table des acronymes

BH	*Benjamini-Hochberg*
DRS	*Data Reduction Software*
EQM	*Erreur Quadratique Moyenne*
FDP	*False Discovery Proportion*
FDR	*False Discovery Rate*
FSF	*Field Spread Function*
FWER	*Family Wise Error Rate*
FWHM	*Full Width at Half Maximum*
HST	*Hubble Space Telescope*
IFU	*Integral Field Unit*
LSF	*Line Spread Function*
MCO	*Moindres Carrés Ordinaires*
MUSE	*Multi-Unit Spectroscopic Explorer*
PSF	*Point Spread Function*
RSB	*Rapport Signal à Bruit*
TSE	*Translatées Linéairement Espacées*
VLT	*Very Large Telescope*



Introduction

L'astronome Carl Sagan disait : "nous sommes poussières d'étoiles". En effet chacun de nos atomes complexes a été forgé au coeur d'une étoile aujourd'hui disparue. Découvrir l'Univers lointain c'est donc partir à la recherche de nos origines et les astronomes n'ont ainsi eu de cesse de développer des instruments de plus en plus sensibles pour sonder cet Univers lointain. A ce titre, le spectro-imageur MUSE (*multi unit Spectroscopic explorer*), mis en place à l'observatoire du VLT [2] en 2014, a révolutionné la spectroscopie astrophysique en permettant l'analyse simultanée d'un grand nombre d'objets célestes. Les données produites par MUSE forment en effet un cube (dit hyperspectral), et peuvent être vues comme un empilement d'images, chacune à une longueur d'onde précise, mais aussi comme un ensemble de spectres très finement échantillonnés, permettant de discriminer la signature de différents composés chimiques émis par ces corps célestes. Pour de tels volumes de données, il ne peut être envisagé de tester chaque pixel, ou chaque spectre, pour savoir si un objet intéressant est présent ou non. Il est donc nécessaire de mettre en place des méthodes de détection et d'analyse afin de traiter ce flot massif de données. Ces méthodes doivent de plus être adaptées aux spécificités de ces données astrophysiques, qui possèdent des particularités très fortes par rapport aux données classiquement étudiées sur Terre (très faibles rapports signaux-à-bruit, impossibilité d'établir des vérités-terrains). On voit donc poindre le besoin de combiner les compétences issues de l'expertise astrophysique et de la science des données. C'est dans ce contexte que se situent mes travaux de thèse, exposés dans le présent manuscrit.

Objectifs et contexte de la thèse

Cette thèse s'inscrit dans le cadre d'une collaboration forte entre le Centre de Recherche Astrophysique de Lyon (CRAL), plus particulièrement le projet ERC 339659-MUSICOS (MUSE Imaging of the COSmic web), et le Gipsa-lab. Parallèlement à cette thèse plusieurs autres thèses et post-doctorats ont été lancés au sein du projet MUSICOS, permettant des échanges soutenus avec le laboratoire ICube de Strasbourg et le laboratoire Lagrange de l'OCA (Observatoire de la Côte d'Azur) de Nice.

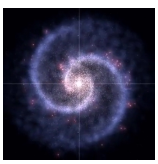
Au titre de cette collaboration pluridisciplinaire, cette thèse devait remplir le double objectif du développement de nouvelles méthodologies de traitement du signal et de leur implémentation et application dans le contexte des données issues de l'instrument MUSE. Cela se traduit notamment par des contraintes fortes du point de vue méthodologique. Les méthodes développées dans cette thèse cherchent ainsi à :

- s'interfacer avec les outils déjà développés au sein du consortium MUSE,
- être utilisables de façon routinière par la communauté astrophysique.

Les méthodes implémentées doivent donc être relativement rapides (afin de permettre une analyse aisée des données) et les plus robustes possibles sur données réelles.

Du point de vue applicatif, le but premier de la thèse était le développement d'outils de détection des structures dans l'environnement proche (circum-galactique) et plus lointain (inter-galactique) des galaxies dans des champs très profonds d'observation de l'Univers lointain issus

2. http ://www.eso.org/public/teles-instr/paranal-observatory/vlt/



de l'instrument MUSE. Une problématique complémentaire a toutefois rapidement vu le jour : ces champs profonds observés par un instrument au sol comme MUSE posent d'importants défis de mélange spectral des sources étudiées. Ainsi cette thèse a finalement cherché à apporter des solutions concrètes à ces deux problématiques issues de l'étude des données astrophysiques MUSE :

- démélange spectral de sources dans un champ profond MUSE,
- détection des structures étendues circum-galactiques.

Ce double objectif se retrouve dans la structuration du présent document, décrite ci-après.

Organisation du document

Le manuscrit est organisé en deux grandes parties pouvant être lues de façon indépendante et dans l'ordre de préférence du lecteur. Elles sont précédées par un premier chapitre, intitulé *Contexte astrophysique*, décrivant le cadre applicatif général de cette thèse. Ce premier chapitre présente notamment l'instrument MUSE et les données que celui-ci fournit, et expose brièvement les deux principales problématiques.

La première partie est intitulée *Démélange spectral de sources* et traite le problème de démélange dans les champs profonds MUSE. Elle est découpée en trois chapitres, suivis d'un bilan et de perspectives Le premier chapitre expose en détail le besoin d'une méthode de démélange adaptée aux données MUSE, par rapport à l'état de l'art. Le deuxième chapitre présente la méthode développée pour réaliser le démélange spectral des sources. Cette méthode s'appuie sur l'utilisation des données à haute résolution spatiale du télescope spatial Hubble. Une première version non régularisée est testée sur données simulées, montrant ses limites et poussant à développer des régularisations adaptées. Le troisième chapitre de cette partie détaille l'application de cette méthode sur les données MUSE du champ ultra-profond UDF.

La seconde partie, intitulée *Détection de sources étendues*, présente le développement d'une méthode de détection des structures étendues environnant les galaxies. Les travaux exposés dans cette partie ont notamment fait l'objet d'un article de journal (Bacher et al. 2017c) et de deux conférences internationales (Bacher et al. 2016a,Bacher et al. 2017b). Cette deuxième partie suit une structure similaire à la première, articulée en trois chapitres et un bilan. Le premier chapitre décrit la problématique de la détection de sources étendues et expose les différentes approches de l'état de l'art. Dans le deuxième chapitre, la méthode proposée est décrite. Cette méthode est fondée sur un test d'hypothèses de type max-test à l'aide d'un dictionnaire et un apprentissage des statistiques de test sur les données. Cette méthode est ensuite étendue pour prendre en compte un a priori de connexité spatiale. Le troisième chapitre expose les résultats obtenus sur les données MUSE du champ ultra-profond UDF, et décrit les pré-traitements mis en place pour cette application. Enfin le bilan et les perspectives autour de cette problématique de détection sont exposés, suivi d'une conclusion générale.

Plusieurs annexes accompagnent ces travaux. L'annexe A présente les preuves des différentes propositions de la partie II, sorties du corps du texte pour en faciliter la lecture. Une méthode de pré-traitement des données a également été développée et présentée en annexe B. Par ailleurs, des travaux exploratoires sur l'étude d'outils de classification pour la détection des sources étendues sont présentés en annexe C.

Publications associées

Les travaux présentés dans cette thèse ont donné lieu à plusieurs publications dans des revues et des conférences à comité de lecture, récapitulées ci-dessous.

Article de journal

[Bacher et al. 2017c] : Raphael Bacher, Céline Meillier, Florent Chatelain et Olivier Michel (2017c). « Robust Control of Varying Weak Hyperspectral Target Detection With Sparse Nonnegative Representation ». In : *IEEE Transactions on Signal Processing* 65.13, p. 3538–3550.
Cet article paru dans *Transactions on Signal Processing, IEEE* détaille la méthode proposée de détection de sources étendues. Il expose la majeure partie des travaux présentés dans la partie II, *Détection de sources étendues*, notamment la construction du test d'hypothèses et des statistiques de test associées, ainsi que leur calibration sous l'hypothèse nulle par un apprentissage sur les données.

La valorisation des travaux dans le domaine applicatif astrophysique, notamment concernant l'aspect démélange, est en préparation, via une publication à venir dans une revue comme *Astronomy and Astrophysics*.

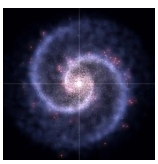
Conférences nationales et internationales

[Bacher et al. 2016a] : Raphael Bacher, Florent Chatelain et Olivier Michel (2016a). « An adaptive robust regression method : application to galaxy spectrum baseline estimation ». In : *IEEE International Conference on Accoustic, Speech and Signal Processing 2016*.
Cet article, présenté sous forme de poster à la conférence IEEE ICASSP 2016, expose les travaux autour de l'étape de pré-traitement d'estimation robuste des lignes de base des spectres. Cette méthode est présentée en détail dans l'annexe B.

[Bacher et al. 2016b] : Raphael Bacher, Florent Chatelain et Olivier Michel (2016b). « Source Halo Advanced Detection and Estimation : une méthode de détection du Circum-Galactic Medium ». In : *Colloque du Groupe Hyperspectral de la Société Française de Photogrammétrie et de Télédétection (SFPT) 2016*.
Une présentation orale des problématiques des données hyperspectrales MUSE a également été faite au colloque 2016 du groupe de recherche hyperspectral de la SFPT.

[Bacher et al. 2017b] : Raphael Bacher, Florent Chatelain et Olivier Michel (2017b). « Global error control procedure for spatially structured targets ». In : *IEEE European Signal Processing Conference 2017*.
Cet article, présenté sous forme de poster à la conférence IEEE EUSIPCO 2017, prolonge les travaux de [Bacher et al. 2017c] en proposant une nouvelle méthode de détection prenant en compte un *a priori* de connexité sur la cible recherchée.

[Bacher et al. 2017a] : Raphael Bacher, Florent Chatelain et Olivier Michel (2017a). « Détection de cibles spatialement structurées sous contrôle global d'erreur ». In : *Colloque du Groupe de recherche et d'étude de traitement du signal et des images (GRETSI) 2017*.
Cet article, présenté sous forme de poster au colloque GRETSI 2017, expose à la communauté française l'approche de détection de sources étendues connexes également développée dans [Bacher et al. 2017b].



Au cours de cette thèse, une contribution a également été apportée à d'autres travaux dans le cadre de l'application MUSE, notamment via une publication dans *Astronomy and Astrophysics* (Meillier et al. 2016), et une communication au colloque GRETSI 2017 (Meillier et al. 2017). Les travaux préliminaires sur les approches par classification ont également fait l'objet d'un chapitre d'ouvrage suite à l'école d'été Basmati (Bacher et al. 2016c).

Chapitre 1

Contexte astrophysique

Sommaire

Ce premier chapitre présente le contexte général de cette thèse qui a pour but de développer des méthodes facilitant l'analyse astrophysique des données MUSE. Pour cela, nous commençons par présenter brièvement le projet MUSE, ainsi que le fonctionnement de l'instrument et le type de données produites. Nous présentons ensuite quels sont les objectifs astrophysiques de l'analyse des champs profonds obtenus par MUSE, comme par exemple l'Ultra Deep Field. Enfin nous exposons les deux principales problématiques abordées dans cette thèse, à savoir le démélange spectral et la détection de sources étendues dans ces champs profonds.

1.1 Le projet MUSE

1.1.1 Le consortium

L'instrument MUSE (*multi-unit spectroscopic explorer*) résulte de nombreuses années de travaux et de la collaboration d'un certain nombre de laboratoires au sein d'un consortium européen [1]. Le consortium MUSE est piloté par le Centre de Recherche Astrophysique de Lyon (CRAL : INSU-CNRS/Université Claude Bernard-Lyon 1/ENS Lyon), sous la direction de Roland Bacon, et comprend les centres de recherche suivants : l'Observatoire européen Austral (ESO), le Leiden Observatory (NOVA - Hollande), l'Institut de recherche en astrophysique et planétologie (IRAP - INSU-CNRS/Université Paul Sabatier ; Observatoire Midi-Pyrénées), l'Institut für Astrophysik (Georg-August University of Göttingen - Allemagne), l'Insitute for Astronomy à ETH, Zurich (Suisse) et l'Astrophysikalisches Institut Potsdam (Allemagne). En plus du consortium, des collaborations fortes, notamment pour l'apport de méthodes de traitement des données, ont été établies avec le Gipsa-lab (Grenoble), iCube (Université de Strasbourg) et le laboratoire Lagrange (Observatoire de la Côte d'Azur, Nice).

1. http://muse.univ-lyon1.fr/spip.php?article97

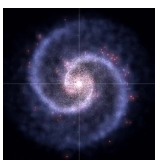


FIGURE 1.1 – Photographie de l'instrument MUSE (à gauche) installé au foyer de l'UT4 (au centre). Photo disponible sur le site de l'ESO, crédit : Eric Le Roux/University Claude Bernard Lyon 1/CNRS/ESO

1.1.2 Instrument

MUSE est un instrument dit de la deuxième génération des appareils installés au VLT (Very Large Telescope), site d'observation de l'ESO (European Southern Obervatory). Le VLT se compose de quatre télescopes primaires et de quatre télescopes secondaires, au foyer desquels sont installés les différents instruments. MUSE se situe au foyer de l'UT4, le quatrième télescope primaire du VLT, qui possède un miroir primaire de 8m de diamètre. On peut voir sur la figure 1.1 une vue de l'instrument installé au foyer du télescope du VLT.

MUSE est un spectrographe intégral de champ (ou spectrographe 3D), c'est-à-dire qu'il fournit non seulement une image du champ observé, mais aussi un spectre pour chacun des pixels de cette image. Contrairement aux imageurs classiques qui intègrent le signal sur toute leur bande spectrale d'observation, MUSE conserve l'information spectrale et permet ainsi d'obtenir une carte du ciel en trois dimensions.

Pour cela MUSE s'appuie sur 24 modules identiques, des *Integral Field Unit* ou IFU, qui contiennent chacun un spectrographe. La lumière arrivant sur l'instrument est redirigée sur un dérotateur de champ qui compense l'effet de la rotation de la Terre sur les observations. La lumière est alors envoyée sur un découpeur de champ qui sépare le champ observé en 24 sous-champs qui sont envoyés chacun sur un des 24 IFU de l'instrument MUSE. Chaque IFU comporte un nouveau découpeur de champ qui découpe le sous-champ en 48 tranches qui sont ensuite envoyées dans le spectrographe. Ce spectrographe sépare la lumière en longueurs d'onde (bandes étroites de 0,125 nm de largeur) et le résultat est enregistré par un capteur CDD de 4000×4000 pixels.

Les données ainsi collectées sont finalement réorganisées sous la forme d'un cube hyperspectral à l'aide d'un système de réduction de données (*Data Reduction Software* ou DRS), décrit brièvement en partie 1.2.2. Grâce à ces jeux de miroirs complexes, MUSE combine ainsi de façon unique un large champ de vue spatial, échantillonné à une bonne résolution spatiale, et une large gamme spectrale observée, à haute résolution spectrale.

1.2 Le champ Ultra Deep Field (UDF)

1.2.1 Description

Un des objectifs de MUSE est pouvoir réaliser des observations de l'Univers extrêmement lointain. On parle alors de champs profonds.

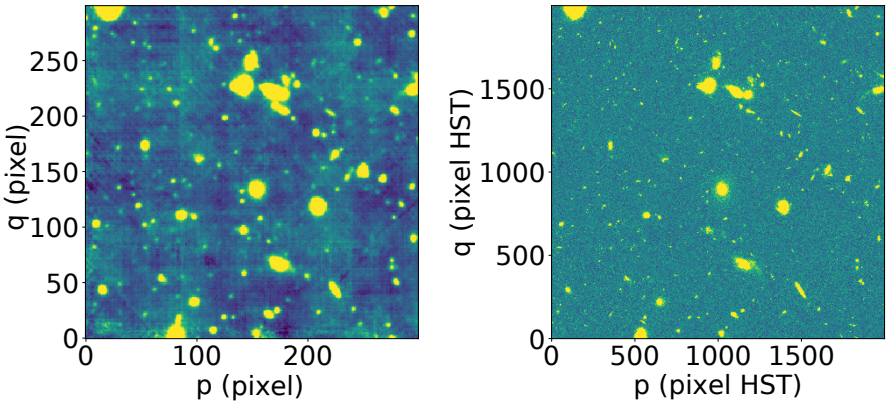


FIGURE 1.2 – A gauche, le champ central de l'UDF observé par MUSE (résolution pixélique de 0.2 arcsec), et son équivalent HST (résolution pixélique de 0.03 arcsec) à droite. Notons que l'image MUSE affichée résulte en fait de la somme des 3600 feuillets spectraux (on parle alors d'*image blanche*).

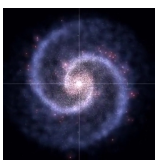
Un des champs les plus profonds jamais réalisés est l'*Ultra Deep Field* ou UDF. Il a été d'abord observé par le télescope spatial Hubble (*Hubble Space Telescope* ou HST) entre septembre 2003 et janvier 2004 (totalisant près de 1 million de secondes de temps d'exposition réparties en séquences typiques de 20 minutes). Ce champ contient près de 10000 galaxies, remontant jusqu'à environ 13 milliards d'années (entre 400 et 800 millions d'années après le Big Bang). Il se situe dans la constellation australe du Fourneau dans une région pauvre en étoiles brillantes (afin de favoriser l'observation d'objets lointains peu lumineux) et s'étend sur environ [2] 9 arcmin^2. L'observation d'un tel champ par MUSE se fait sur plusieurs mois et des dizaines de nuit afin d'acquérir à la fois les poses scientifiques et les poses nécessaires à la calibration des données. Afin de couvrir un champ équivalent à celui d'Hubble, le champ UDF de MUSE est le résultat d'une mosaïque de 3 par 3 champs individuels. La zone centrale a également été observée spécifiquement par ailleurs, permettant la création d'un dixième champ plus profond que l'ensemble de la mosaïque. Ce champ central est illustré par la figure 1.2 et le positionnement de l'ensemble des champs MUSE par rapport au champ HST est présenté sur la figure 1.3.

1.2.2 Réduction des données

A partir de l'acquisition d'observation chaque nuit, un certain nombre d'opérations sont effectuées pour obtenir le cube de données tel qu'exploité par la suite. Ces étapes sont détaillées précisément dans [BACON et al. 2015 ; BACON et al. 2017] et dans [WEILBACHER et al. 2012], nous rappelons ici brièvement la chaîne de traitement :

1. Acquisition de poses individuelles et de données de calibration (biais, dark, offset). Les poses sont stockées et manipulées sous la forme de *pixtables*, tableaux de données comportant également l'information de position spatiale et spectrale.
2. Corrections à l'aide de ces données de calibration.
3. Correction astrométrique (calibration du flux sur toutes les poses)

2. soit l'équivalent en diamètre d'un dixième du diamètre de la pleine lune.



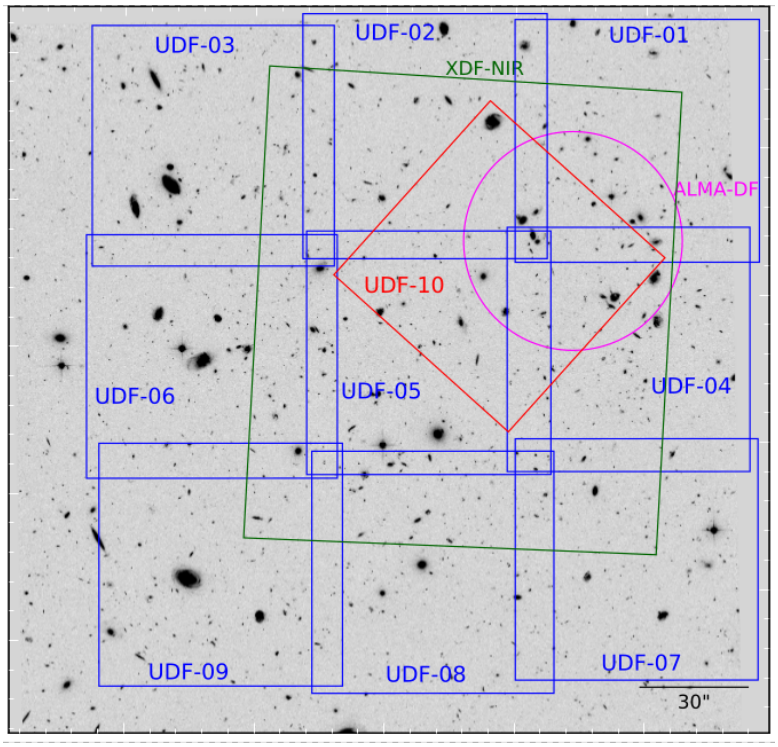


FIGURE 1.3 – Le champ UDF du HST et la mosaïque des 9 champs MUSE correspondants (en bleu), plus le dixième champ central (en rouge). Le champ XDF-NIR (en vert) indique la zone observée par HST en proche infra-rouge, et la zone magenta indique le champ observé par l'instrument ALMA. Extrait de [BACON et al. 2017].

4. Réduction des écarts systématiques entre capteurs en ramenant les observations de chaque capteur autour de la même valeur médiane.

5. Interpolation par *drizzling* en trois dimensions (algorithme adapté pour les trois dimensions de MUSE à partir des travaux de FRUCHTER 2009) qui permet d'aligner les différentes poses individuelles sur la même grille de pixels que le cube final.

6. Soustraction du ciel sur chaque pose individuelle : une méthode appelée ZAP a été développée SOTO et al. 2016 au sein du consortium pour répondre à ce problème. Cette méthode s'appuie sur une analyse en composante principale (ACP).

7. Fusion des poses individuelles à l'aide d'une méthode de σ-clipping (permettant de rejeter les mesures aberrantes).

Si d'un point de vue astrophysique tous ces traitements visent à supprimer des effets physiques indésirables, d'un point de vue traitement du signal certaines opérations de la chaîne de réduction des données induisent des effets non négligeables sur les données. Notamment l'opération de *drizzling* en trois dimensions induit une certaine corrélation (spatiale et spectrale) des pixels avec leurs proches voisins, difficile à modéliser (du fait de la taille des données).

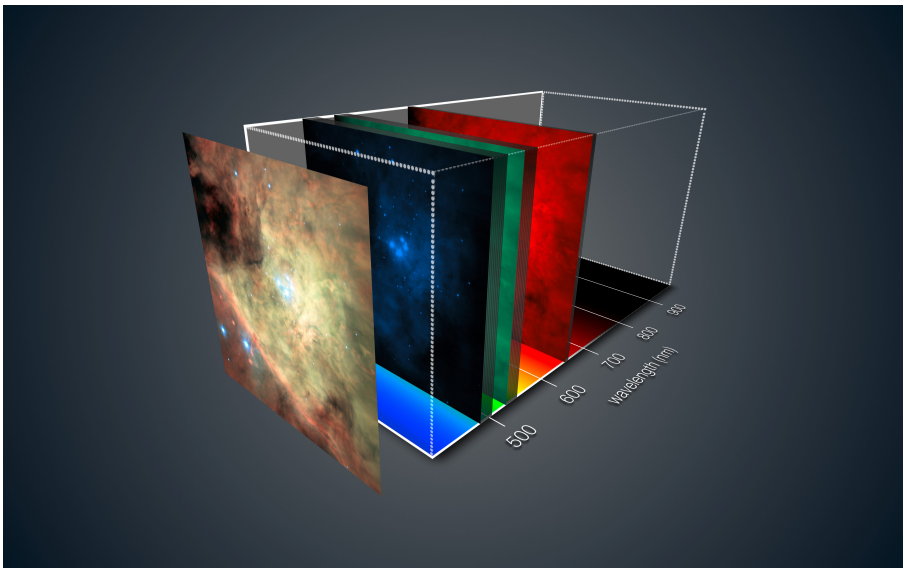


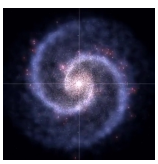
FIGURE 1.4 – Un cube hyperspectral MUSE peut être vu comme une superposition d'images prises dans près de 4000 longueurs d'onde dans le domaine visible et proche infrarouge. Cette décomposition des couleurs permet notamment de révéler la composition chimique et les propriétés physiques des objets étudiés. Ici une mosaïque de plusieurs observations de la nébuleuse d'Orion par MUSE début 2014.
Credit : ESO/MUSE consortium/R. Bacon/L. Calçada

1.2.3 Modélisation des données hyperspectrales

1.2.3.1 Format des données et notations

Les données finales sont fournies sous la forme d'un cube de données, composé d'une dimension spectrale et de deux dimensions spatiales. On parle alors de cube hyperspectral. Les dimensions d'un cube final sont notées $p \times q \times \lambda$ avec $p \approx q \approx 300$ pixels (de 0.2 arcsec de côté) et $\lambda = 3600$ longueurs d'onde dans le domaine visible et proche infrarouge (entre 475nm et 950nm). Chaque cube peut être interprété comme une superposition d'images prises dans ces différentes longueurs d'onde, comme illustrée sur la figure 1.4, ou comme un ensemble de spectres organisés spatialement (figure 1.5). Notons que cette deuxième interprétation (ensemble de spectres) correspond à la réalité du processus d'acquisition, MUSE ne fonctionnant pas par superposition de filtres spectraux mais bien par analyse spectroscopique de chaque partie du champ.

On désigne par *voxel* une valeur élémentaire de ce cube de données, repéré par un triplet *(p,q,λ)*. On parlera indifféremment de pixel ou de spectre pour désigner un vecteur spectral repéré par un tuple *(p,q)*. On appelle feuillet l'ensemble des voxels à une longueur d'onde λ donnée. Dans la plupart des traitements, le cube de données sera vu comme une matrice, de dimensions $\lambda \times n$ avec $n = pq$, où chaque ligne correspond à un feuillet vectorisé. On parle d'*image blanche* pour désigner l'image résultante de la somme de tous les feuillets, et d'image en *bande étroite* pour désigner une somme de quelques feuillets autour d'une longueur d'onde donnée.



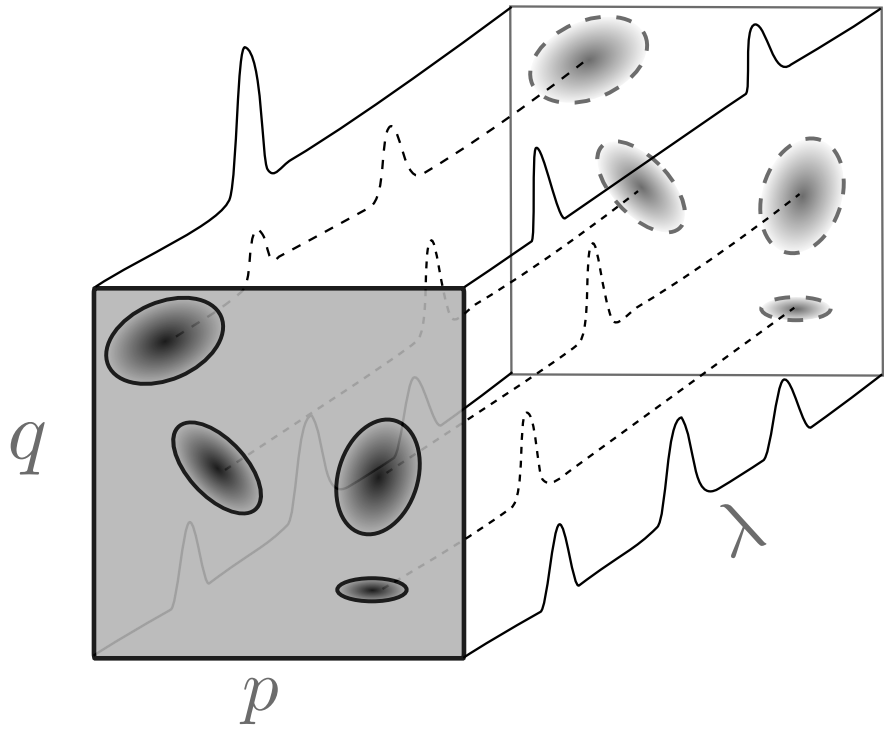


FIGURE 1.5 – Un cube hyperspectral MUSE peut également être vu comme un agencement spatial de différents spectres. Ces spectres peuvent être associés aux objets (galaxies, étoiles) présents dans le champ.

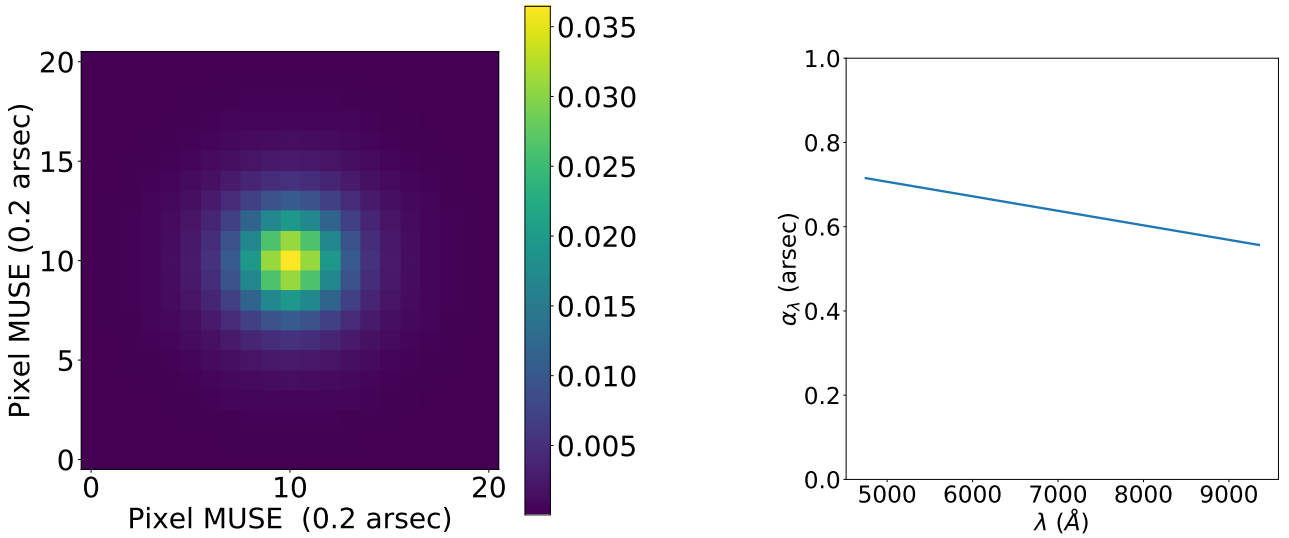
1.2.3.2 PSF

Comme tout instrument de mesure, MUSE possède une réponse impulsionnelle qui va venir déformer le signal observé. On appelle cette réponse la fonction d'étalement du point (*Point Spread Function* ou PSF). Dans le cadre d'instrument d'observation céleste depuis le sol, cette PSF est en fait la combinaison des effets de l'instrument lui-même, du télescope et des turbulences atmosphériques. Cette PSF est une fonction à 3 dimensions (2 dimensions spatiales et une spectrale) complexe à étudier.

Un certain nombre de travaux ont déjà été menés sur l'étude de la PSF de MUSE (CARFANTAN 2014 ; VILLENEUVE et al. 2011 ; VILLENEUVE 2012). Il en ressort que la PSF peut être décomposée en la multiplication d'une composante spatiale (*Field Spread Function* ou FSF) et d'une composante spectrale (*Line Spread Function* ou LSF). Ces deux composantes peuvent être supposées invariantes spatialement mais dépendent de la longueur d'onde. La LSF dépend uniquement de l'instrument et est donc estimée directement sur les données de calibration de l'instrument via une modélisation par une gaussienne tronquée dont la variance varie très légèrement avec la longueur d'onde. Son influence est négligeable dans le cadre nos applications car les objets célestes étudiés ont des caractéristiques spectrales grandes devant la LSF. Le profil de la LSF, pour des longueurs d'onde au début et à la fin du domaine spectral de MUSE, est présenté sur la figure 1.7.

La FSF est quant à elle principalement due aux turbulences de l'atmosphère et doit ainsi être estimé pour chaque observation.

En l'absence d'optique adaptative la FSF est modélisée par une fonction de Moffat (MOFFAT 1969), modèle régulièrement utilisé pour modéliser la FSF spatiale en astrophysique, voir par



(a) Profil spatial de la FSF pour le champ central UDF à 6000Å

(b) évolution du paramètre d'échelle α_λ en fonction de la longueur d'onde

FIGURE 1.6 – Profil spatial de la FSF pour $\lambda \approx 6000$Å et évolution du paramètre d'échelle α_λ en fonction de la longueur d'onde.

exemple TRUJILLO et al. 2001. Elle a pour expression :

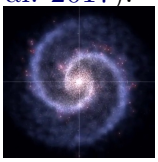
$$F_\lambda(p,q) = \frac{\beta - 1}{\pi \alpha^2} \left(1 + \frac{p^2 + q^2}{\alpha_\lambda^2}\right)^{-\beta_\lambda},$$

où α_λ et β_λ sont, respectivement, le paramètre d'échelle et le paramètre de forme de la fonction Moffat. Ces paramètres dépendent de la longueur d'onde λ, car la turbulence impacte plus fortement les petites longueurs d'onde. La largeur à mi-hauteur, abrégée FWHM, pour *Full-Width at Half Maximum* s'obtient alors par : $\mathrm{FWHM}_\lambda = 2\alpha_\lambda\sqrt{2^{1/\beta_\lambda} - 1}$. Des travaux au sein du consortium permettent d'estimer sur un champ donné les valeurs de α_λ (évoluant linéairement en fonction de la longueur d'onde) et β_λ (supposé constant en longueur d'onde, voir [VILLENEUVE et al. 2011]). Un exemple de profil de FSF dans l'UDF est présenté dans la figure 1.6 ainsi que l'évolution du paramètre d'échelle estimé dans le champ central de l'UDF.

Nous supposerons la LSF et la FSF connues par la suite.

1.2.3.3 Bruit

Les données produites par MUSE sont extrêmement bruitées au regard des objets auxquels nous allons nous intéresser. Ce bruit est le résultat de plusieurs sources de bruit qui se mélangent, notamment les émissions parasites de l'atmosphère, et les bruits de mesure au niveau des capteurs (bruit thermique notamment). La modélisation de toutes ces contributions de bruit adoptée dans le consortium MUSE repose sur un modèle de bruit additif gaussien dont la moyenne est constante spatialement, mais dont la variance varie dans les trois dimensions du cube. Cette variance peut notamment varier fortement d'une bande spectrale à l'autre suivant la présence ou non d'une raie du ciel (estimée puis soustraite lors des pré-traitements). Un cube de variance est donc fourni avec chaque cube de données, en propageant une estimation de la variance en chaque point (voxel) du cube depuis les capteurs (voir BACON et al. 2017).



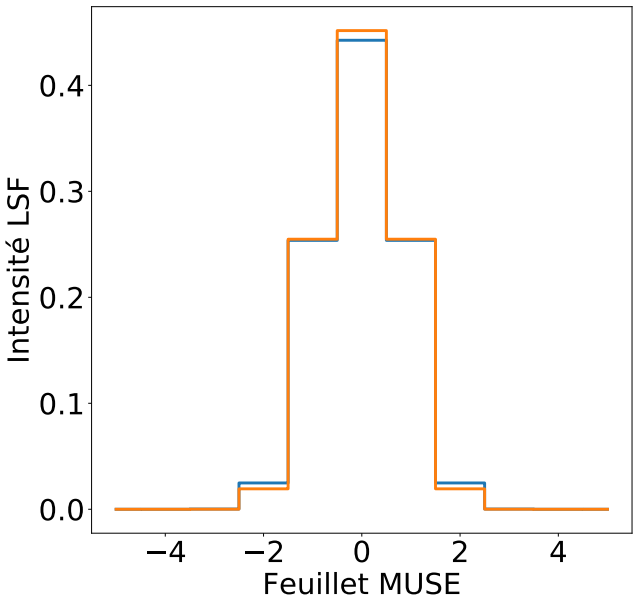


FIGURE 1.7 – Profil de la composante spectrale LSF (après échantillonnage à la résolution spectrale MUSE) à 5000Å (bleu) et 9000Å (orange).

Toutefois les différentes étapes du processus de réduction de données créent une structure de dépendance entre voxels et la taille des données rend très compliqué toute modélisation de ces dépendances. Ces corrélations sont toutefois de faible portée car elles proviennent principalement de l'étape de *drizzling* lors de la réduction des données. L'étude de ce bruit, menée notamment dans COURBOT 2017, permet de montrer que si les poses individuelles d'une observation MUSE contiennent un bruit modélisable par des distributions de Student, le cube final, obtenu en intégrant ces différentes poses, possède lui un bruit quasi-gaussien, mais toujours corrélé. Cette quasi-gaussianité, tout comme la loi de Student, signifie notamment une symétrie de la distribution de ce bruit, qui sera utilisée dans les travaux présentés dans le manuscrit.

1.3 Analyse des champs profonds MUSE

Cette technique de spectroscopie intégrale de champ permet aux astronomes d'étudier les propriétés spectrales des différentes parties d'un objet résolu telle qu'une galaxie afin par exemple d'observer sa rotation et d'en déduire sa masse. Elle permet également de déterminer la composition chimique ainsi que les propriétés physiques des différentes régions de l'objet étudié. Cela permet de plus à MUSE d'être particulièrement sensible aux objets qui émettent la majeure partie de leur énergie à certaines longueurs d'onde bien spécifiques. Les galaxies du jeune Univers présentent typiquement ce genre de spectre car elles contiennent de l'hydrogène gazeux qui possède des raies d'émission très brillantes.

Travaux précédents sur les champs profonds : détection de sources ponctuelles

Les données MUSE, en tant que données hyperspectrales dans un cadre astrophysique, sont d'un genre nouveau. En effet il s'agit des premières données hyperspectrales de cette ampleur

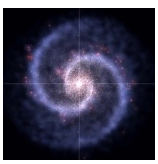
à destination de la communauté astrophysique, ce qui rend l'utilisation des outils d'analyse classiquement utilisés par cette communauté peu adapté à ces données. De façon similaire, les outils développés classiquement pour l'analyse de données hyperspectrales se placent en grande majorité dans le domaine de la télédétection terrestre, domaine aux caractéristiques très différentes de l'observation astrophysique (rapport signaux-à-bruit bien plus favorables, peu de composantes spectrales, moins de bandes spectrales).

De nombreux travaux ont donc été développés récemment autour du projet MUSE afin d'aider à l'analyse des champs profonds MUSE mais la plupart se sont concentrés sur la détection de sources quasi-ponctuelles (à la PSF près). En effet les méthodes de détection classiques d'objets astrophysiques, tel Sextractor (Bertin et Arnouts 1996) ne sont pas aisément adaptables aux données 3D de MUSE, et nécessitent le plus souvent des réglages fins de leurs paramètres afin d'éviter de trop nombreuses fausses alarmes. On peut ainsi citer notamment le développement d'une méthode de détection s'appuyant sur une approche par processus ponctuel marqué et une modélisation objet (Meillier et al. 2016), ainsi que des approches par test d'hypothèses GLR (*generalized likelihood ratio*) [Paris et al. 2013]. D'autres approches ont également tenté d'approcher ce problème à l'aide de modélisations parcimonieuses des signatures recherchées (Bourguignon et al. 2012). Ces différents travaux permettent ainsi d'obtenir les nombreuses galaxies d'un champ, qui sont assimilables pour une grande partie à des sources ponctuelles et restent relativement lumineuses par rapport aux structures extra-galactiques.

Dans cette thèse, nous considérons la détection de ces galaxies comme acquise et nous nous concentrons sur deux nouvelles problématiques, le démélange spectral des galaxies et la détection de sources étendues dans l'environnement de ces galaxies.

1.4 Démélange spectral de galaxies

L'étude par MUSE d'un champ profond comme l'UDF permet de caractériser les propriétés physico-chimiques d'un grand nombre de galaxies à partir de leur spectre. Il est donc nécessaire de pouvoir obtenir les caractéristiques spectrales de chaque galaxie. Or, si la résolution spectrale de MUSE est de grande qualité, sa résolution spatiale est limitée, du fait notamment de la présence au sol de l'instrument. En effet les turbulences atmosphériques dégradent le signal parvenant aux capteurs. Cette dégradation est modélisée dans la PSF de MUSE, comme une convolution spatiale appliquée au signal émis par les galaxies. Cette dégradation spatiale crée une situation de mélange spectral, chaque pixel pouvant contenir les contributions spectrales de plusieurs objets différents. L'objectif de l'opération de démélange est de pouvoir retrouver la signature spectrale de chaque source. La problématique de démélange est un sujet couramment étudié pour des données hyperspectrales, le plus souvent dans le cadre de la télédétection (observation de la Terre à l'aide d'un capteur aéroporté ou satellisé). Les données MUSE présentent toutefois des spécificités importantes par rapport aux données de télédétection, rendant l'utilisation des méthodes classiques peu pertinente. De plus, le champ UDF a également été imagé par HST, qui possède une résolution spatiale bien meilleure que MUSE (étant notamment non soumis aux turbulences atmosphériques). Nous proposons donc dans la partie I une méthode s'appuyant sur les données HST pour estimer les spectres des sources présentes dans un champ MUSE.



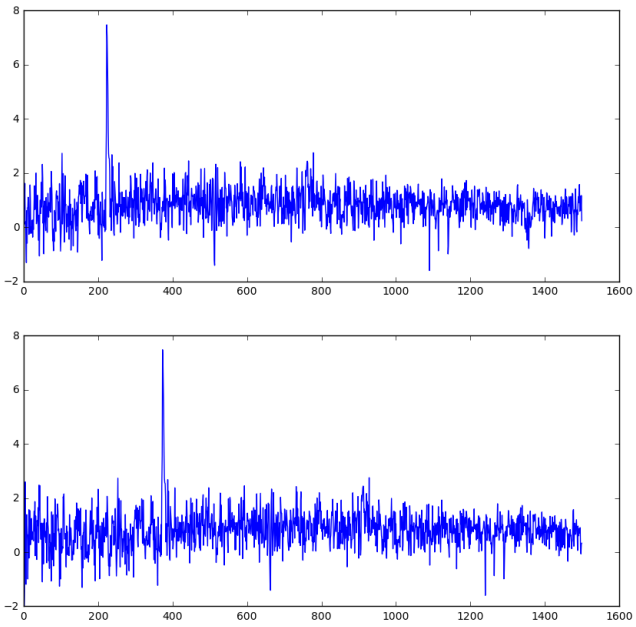


FIGURE 1.8 – Effet du redshift. En haut : spectre au repos (par ex. émis en laboratoire) ; en bas : spectre d'un objet lointain observé sur Terre.

1.5 Détection de sources étendues

Un des grands défis actuels en astronomie est la compréhension de l'interaction des galaxies avec leur environnement proche, notamment via les flux de matière gazeuse qui alimentent les galaxies ou s'en échappent. L'étude de ces flux permet en effet de mieux comprendre la formation stellaire au sein des galaxies et la propagation des éléments chimiques lourds (créés au sein des étoiles). Cet environnement proche est appelé le Circum Galactic Medium ou CGM.

Il est principalement composé de gaz d'hydrogène. Cet hydrogène émet notamment très fortement dans l'ultraviolet (à 121.5 nm) une raie d'émission, la raie Lyman-α. C'est cette raie d'émission que l'on va chercher à détecter en tant que marqueur du CGM.

Toutefois, lorsqu'on s'intéresse à des objets célestes lointains, le phénomène du décalage vers le rouge ou *redshift* (voir par exemple LIDDLE 2015) doit être pris en compte. Du fait de l'expansion de l'Univers, le spectre d'un objet observé depuis la Terre se verra décalé en longueur d'onde selon la formule :

$$1 + z = \frac{\lambda_{obs}}{\lambda_0}$$

où z est la valeur du redshift, λ_{obs} la longueur d'onde observée de la raie et λ_0 la longueur d'onde de la raie au moment de son émission (longueur d'onde au repos).

Un redshift nul correspond ainsi à un objet très proche, et un redshift élevé correspond à un objet lointain. Trois phénomènes contribuent au redshift : l'expansion de l'Univers, le mouvement particulier de l'objet et la cinématique de l'objet. Toutefois, pour les objets lointains, les deux derniers phénomènes sont négligeables par rapport au premier.

On peut voir sur la figure 1.8 l'effet de ce redshift sur un spectre d'émission.

Autrement dit, pour des objets suffisamment lointains (autour de 12 milliards d'années-lumière), l'émission Lyman-α, émise à 121.5 nm, devient observable dans le domaine spectral de MUSE (entre 475 et 935nm) et on ne peut connaître a priori la position spectrale de la raie.

Cela implique également que des sources spatialement étendues, tels que des halos de CGM peuvent posséder une certaine variabilité spectrale.

Par ailleurs le CGM émet de façon très faible par rapport aux autres sources présentes dans des données MUSE (galaxies et étoiles). Il est donc nécessaire de pouvoir se prémunir contre les contributions de ces sources qui peuvent noyer le signal recherché. Enfin, les différents prétraitements effectués en amont rendent difficile la modélisation du fond de ciel et peuvent donc compliquer la tâche de détection.

On peut donc ramener la problématique de la détection du CGM dans les données MUSE à un problème de détection avec les caractéristiques suivantes :

- recherche d'un signal de faible intensité,
- avec une signature spectrale proche de celle, connue, de la raie Lyman de la galaxie,
- avec une possible variabilité spectrale (principalement un décalage spectral),
- avec une possible pollution des sources environnantes,
- dans un fond difficile à modéliser.

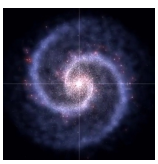
Enfin, le nombre de pixels/spectres à tester étant de grande taille, il apparaît indispensable d'assurer un contrôle global des erreurs le plus robuste possible. Pour résoudre ce problème, nous proposons dans la partie II une méthode de détection fondée sur une approche par test d'hypothèses avec contrôle global du taux d'erreurs.

Résumé

Ces travaux se placent dans le contexte de l'étude des champs profonds hyperspectraux produits par l'instrument d'observation céleste MUSE. Ces données permettent de sonder l'Univers lointain et d'étudier les propriétés physiques et chimiques des premières structures galactiques et extra-galactiques. L'étude de ces champs profonds pose notamment deux problèmes majeures :

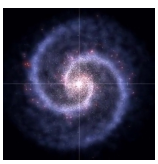
- l'attribution d'une signature spectrale pour chaque source galactique. MUSE étant un instrument au sol, la turbulence atmosphérique dégrade fortement le pouvoir de résolution spatiale de l'instrument, ce qui génère des situations de mélange spectral pour un grand nombre de sources.
- la détection du Circum-Galactic Medium (CGM). Le CGM, milieu gazeux s'étendant autour de certaines galaxies, se caractérise par une signature spatialement diffuse et de faible intensité spectrale.

Nous allons donc désormais exposer en détail les solutions apportées à la problématique de démélange (partie I) et à la détection de sources étendues (partie II).



Première partie

Démélange spectral de sources



Notations et équations - partie I

Notations

- $\boldsymbol{M}$: matrice composée des éléments $m_{i,j}$
- $\boldsymbol{v}$: vecteur
- $\text{diag}(s_j)$: matrice diagonale dont les éléments diagonaux sont les s_j
- $||\boldsymbol{M}||_2 = \sqrt{\sum_{i,j} m_{i,j}^2}$ (norme de Frobenius)
- $\text{Tr}(\boldsymbol{M})$: trace de $\boldsymbol{M}$
- $\boldsymbol{M}^+$: pseudo inverse de $\boldsymbol{M}$
- $\boldsymbol{A} \circ \boldsymbol{b} = \begin{pmatrix} a_{11} \cdot b_1 & \cdots & a_{1n} \cdot b_1 \\ \vdots & \ddots & \vdots \\ a_{m1} \cdot b_m & \cdots & a_{mn} \cdot b_m \end{pmatrix}$. (produit de Hadamard)
- N : nombre de pixels HST
- n : nombre de pixels MUSE
- k : nombre de sources
- λ : nombre de bandes spectrales MUSE
- Λ : nombre de bandes spectrales HST
- $\underset{n\times\lambda}{\boldsymbol{Y}}$: matrice de données issue de MUSE
- $\underset{N\times\Lambda}{\boldsymbol{Z}}$: matrice de données issue de HST
- $\underset{N\times\lambda}{\boldsymbol{X}}$: matrice du champ super-résolu spatialement et spectralement
- $\underset{N\times k}{\boldsymbol{U}}$: matrice d'abondance (résolution HST)
- $\underset{n\times k}{\widetilde{\boldsymbol{U}}}$: matrice d'abondance (résolution MUSE)
- $\underset{k\times\lambda}{\boldsymbol{D}}$: matrice des spectres
- $\underset{n\times N}{\boldsymbol{B}_l}$: matrice de dégradation spatiale pour la bande spectrale l
- $\underset{N\times k}{\boldsymbol{H}}$: matrice de segmentation HST
- $\underset{\lambda\times\Lambda}{\boldsymbol{A}}$: matrice des réponses des filtres HST
- $\mathcal{N}$: bruit d'observation et de mesure
- $s_{max}(\boldsymbol{M}), s_{min}(\boldsymbol{M})$: plus grande et plus petite valeur singulière de $\boldsymbol{M}$
- $\mathcal{M}$: modèle de régression (ensemble de spectres non nuls)
- c : nombre de conditionnement
- f : indice de fidélité d'un spectre
- ic : inter-corrélation entre deux spectres

Equations

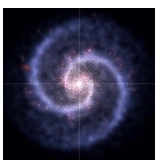
$$\text{Pour } 1 \leq l \leq \lambda, \quad \underset{n\times1}{\boldsymbol{Y}_l} = \underset{n\times N}{\boldsymbol{B}_l} \times \underset{N\times1}{\boldsymbol{X}_l} + \underset{n\times1}{\mathcal{N}} \tag{3.1}$$

$$\underset{N\times\Lambda}{\boldsymbol{Z}} = \underset{N\times\lambda}{\boldsymbol{X}} \times \underset{\lambda\times\Lambda}{\boldsymbol{A}} \tag{3.2}$$

$$\widehat{\boldsymbol{D}}_{i,l} = \underset{\boldsymbol{D}_{i,l}}{\text{argmin}} \, ||\boldsymbol{Y}_l - \widetilde{\boldsymbol{U}}_{i,l}\boldsymbol{D}_{i,l}||_2^2. \tag{3.3}$$

$$\widehat{\boldsymbol{D}}_l = \sum_i a_{i,l}\widehat{\boldsymbol{D}}_{i,l}, \quad 1 \leq l \leq \lambda, \tag{3.4}$$

$$\boldsymbol{U}_i = \underbrace{\boldsymbol{Z}_i}_{N\times1}\underbrace{\boldsymbol{H}}_{N\times k} \tag{3.5}$$



$$\underbrace{\widetilde{\boldsymbol{U}}_{i,l}}_{n\times k} = \underbrace{\boldsymbol{B}_l}_{n\times N}\underbrace{\boldsymbol{U}_i}_{N\times k}, \tag{3.6}$$

$$\widehat{\boldsymbol{D}}_{i,l} = (\widetilde{\boldsymbol{U}}_{i,l}^T\widetilde{\boldsymbol{U}}_{i,l})^{-1}\widetilde{\boldsymbol{U}}_{i,l}^T\boldsymbol{Y}_l. \tag{3.7}$$

$$\underbrace{\mathrm{var}(\widehat{\boldsymbol{D}}_{i,l})}_{k\times k} = (\widetilde{\boldsymbol{U}}_{i,l}^T\widetilde{\boldsymbol{U}}_{i,l})^{-1}\widetilde{\boldsymbol{U}}_{i,l}^T\boldsymbol{\Sigma}_l\widetilde{\boldsymbol{U}}_{i,l}(\widetilde{\boldsymbol{U}}_{i,l}^T\widetilde{\boldsymbol{U}}_{i,l})^{-1}. \tag{3.8}$$

$$c = \frac{s_{max}(\widehat{\boldsymbol{U}})}{s_{min}(\widehat{\boldsymbol{U}})}, \tag{3.9}$$

$$v = \frac{1}{k}\sum_{j=1}^{k}\mathrm{var}(\widetilde{\boldsymbol{d}}_j) \tag{3.10}$$

$$f = \frac{1}{k}\sum_{j=1}^{k}\frac{\langle\widehat{\boldsymbol{d}}_j, \boldsymbol{d}_j\rangle}{||\widehat{\boldsymbol{d}}_j||.||\boldsymbol{d}_j||} \tag{3.11}$$

$$ic = \frac{\langle\widehat{\boldsymbol{d}}_0, \widehat{\boldsymbol{d}}_1\rangle}{||\widehat{\boldsymbol{d}}_0||.||\widehat{\boldsymbol{d}}_1||} \tag{3.12}$$

$$\boldsymbol{y} = \widetilde{\boldsymbol{U}}\boldsymbol{d} + \boldsymbol{\epsilon}, \tag{3.13}$$

$$\widehat{\boldsymbol{d}} = \underset{\boldsymbol{d}}{\mathrm{argmin}}\,||\boldsymbol{y} - \widetilde{\boldsymbol{U}}\boldsymbol{d}||_2^2 + \alpha||\boldsymbol{d}||_2^2, \tag{3.14}$$

$$\widehat{\boldsymbol{d}} = \left(\widetilde{\boldsymbol{U}}^T\widetilde{\boldsymbol{U}} + \alpha\boldsymbol{I}_k\right)^{-1}\widetilde{\boldsymbol{U}}^T\boldsymbol{y}, \tag{3.15}$$

$$\widehat{\boldsymbol{d}} = \underset{\boldsymbol{d}}{\mathrm{argmin}}\,||\boldsymbol{y} - \widetilde{\boldsymbol{U}}\boldsymbol{d}||_2^2 \text{ avec } ||\boldsymbol{d}||_2 \leq t, \tag{3.16}$$

$$\widehat{\boldsymbol{d}} = \underset{\boldsymbol{d}}{\mathrm{argmin}}\,\frac{1}{2}\times||\boldsymbol{y} - \widetilde{\boldsymbol{U}}\boldsymbol{d}||_2^2 + \alpha\times||\boldsymbol{d}||_1, \tag{3.17}$$

$$\frac{1}{2n}\times||\boldsymbol{Y} - \widetilde{\boldsymbol{U}}\boldsymbol{D}||_2^2 + \alpha||\boldsymbol{D}||_{21}, \tag{3.18}$$

$$\mathrm{BIC}(\mathcal{M}) = K\log(n) + \log(\widehat{\sigma_{\mathcal{M}}}^2) \tag{3.19}$$

$$\boldsymbol{D}_0 = \widetilde{\boldsymbol{U}}^T\boldsymbol{Y} \tag{3.20}$$

$$\widetilde{\boldsymbol{d}} = \boldsymbol{d} * \boldsymbol{g}_w \tag{3.21}$$

$$\mathcal{S}_0 = \{T \text{ tel que } \widetilde{d}_T = \max_{T-w\leq t\leq T+w}\widetilde{d}_t\} \tag{3.22}$$

$$\widehat{\sigma}_{\mathrm{MAD}} = \frac{1}{\Phi^{-1}(3/4)}\,\mathrm{MAD}(\widetilde{\boldsymbol{d}}), \tag{3.23}$$

$$\mathcal{S}_1 = \{T \in \mathcal{S}_0 \text{ tel que } \widetilde{d}_T =\leq a\widehat{\sigma}_{\text{MAD}}\} \tag{3.24}$$

$$\mathcal{S}_2 = \{T \in \mathcal{S}_1 \text{ tel que } \forall -1 \leq i \leq 1 \quad d_{T+i} \geq \text{sgn}(d_T)\widehat{\sigma}_{\text{MAD}}\} \tag{3.25}$$

$$w_l = \underset{j}{\text{argmax}} \frac{\boldsymbol{d}_l^T \boldsymbol{g}_j}{||\boldsymbol{d}_l||.||\boldsymbol{g}_j||} \tag{3.26}$$

$$\mathcal{M}_l = \underset{\mathcal{M}}{\text{argmin}}\, \text{BIC}(\mathcal{M}), \tag{3.27}$$

$$\widehat{\boldsymbol{D}}_l^r = \underset{\boldsymbol{D}}{\text{argmin}}\, ||\boldsymbol{Y}_l^r - \widetilde{\boldsymbol{U}}_{\mathcal{M}_l}\boldsymbol{D}||_2^2 \tag{3.28}$$

$$\widehat{\boldsymbol{d}}(\alpha) = \boldsymbol{V}\, \text{diag}\left(\frac{s_j}{s_j^2 + \alpha}\right)\boldsymbol{A}^T\boldsymbol{y} \tag{3.29}$$

$$cv_i = [\boldsymbol{y}_i - (\widetilde{\boldsymbol{U}}\widehat{\boldsymbol{d}}_{-i})_i]^2 \tag{3.30}$$

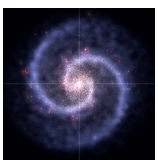
$$GCV = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - (\widetilde{\boldsymbol{U}}\widehat{\boldsymbol{d}})_i}{1 - \text{Tr}(\boldsymbol{S})/n}\right)^2. \tag{3.31}$$

$$\sigma_{CV} = \frac{1}{n\sqrt{n}}\sum_{1\leq i\leq n} cv_i^2, \tag{3.32}$$

$$\alpha_{m+1s} = \max_{\alpha}\{CV(\alpha) \leq CV_m + \sigma_{CV}\} \tag{3.33}$$

$$f_j = \frac{\left\langle \left(\boldsymbol{Y}^c\left(\widehat{\boldsymbol{D}}^c\right)^+\right)_j, \widetilde{\boldsymbol{U}}_j \right\rangle}{||\widetilde{\boldsymbol{U}}_j||_2^2}, \text{ pour } 1 \leq j \leq k. \tag{3.34}$$

$$\min_{\boldsymbol{X}} ||\boldsymbol{Y} - \underset{Data}{\boldsymbol{X}\boldsymbol{B}}||_2^2 + \alpha||\boldsymbol{X} - \underset{Reg.}{\boldsymbol{D}\boldsymbol{U}}||_2^2 \tag{4.1}$$



Chapitre 2

Problématique de démélange

Sommaire

Ce premier chapitre de la partie "Démélange spectral de sources" expose tout d'abord dans la section 2.1 la problématique de mélange spectral dans le cadre de l'application MUSE visée. Un état de l'art des approches de démélange existantes, dans le domaine des données hyperspectrales et dans le cadre plus large de la séparation de sources, est exposé dans la section 2.2. Nous verrons que ces différentes approches ne sont pas adaptées aux spécificités des données MUSE (nombre très important de sources et de bandes spectrales, faible rapport signal-à-bruit), ce qui motive le développement d'une méthode spécifique. Une contrainte forte sur la méthode à développer est son utilisation aisée et régulière par les astronomes sur des données réelles massives, ce qui implique de s'orienter sur une approche rapide, robuste et non-supervisée.

2.1 Contexte

2.1.1 Limitation spatiale de MUSE

Comme détaillé dans le chapitre 1, le spectrographe intégral de champ MUSE permet de produire des cubes hyperspectraux, c'est à dire des données de dimensions $p \times q \times \lambda$ avec $p \approx q \approx 300$ pixels (de 0.2 arcsec de côté) et $\lambda = 3600$ longueurs d'onde, prises dans le domaine visible et proche infrarouge (entre 475nm et 950nm). Chaque cube peut être interprété comme une superposition d'images prises à ces différentes longueurs d'onde ou comme un ensemble de spectres organisés spatialement, comme illustré sur la figure 1.5. Ce type de données permet l'analyse de la composition chimique et des propriétés physiques des sources présentes dans le champ observé.

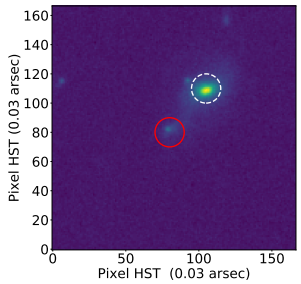
Toutefois, si la résolution spectrale de MUSE est de très grande qualité, l'analyse de ses données est limitée par sa résolution spatiale. En effet, en plus de la réponse impulsionnelle propre à l'instrument, les observations de MUSE sont dégradées par les turbulences de l'atmosphère contrairement aux observations obtenues depuis l'espace. Cette dégradation est prise en compte dans le modèle de la composante spatiale de la fonction d'étalement du point de MUSE (*Field Spread Function* ou FSF. Rappelons en effet que la réponse impulsionnelle de l'instrument MUSE (ou plus précisément de l'ensemble {instrument + atmosphère}) peut être séparée

entre une composante spatiale, la FSF, et une composante spectrale, la *Line Spread Function* ou LSF (voir chapitre 1 pour plus de détails). Les observations subissent donc spatialement une convolution par cette FSF qui possède une largeur à mi-hauteur (*Full Width Half Length* ou FWHM) d'environ 4 pixels MUSE en moyenne, soit environ 0.8 arcsec (notons qu'en pratique cette FSF peut fortement varier d'une pose à l'autre, typiquement entre 0.6 et 1.5 arcsec, en fonction de la turbulence). Aussi, dans les champs profonds, un certain de nombre de sources (galaxies) [1] sont trop proches spatialement les unes des autres pour pouvoir les distinguer à la résolution de MUSE. Cela rend difficile l'analyse spectrale de chacun de ces objets, c'est à dire l'attribution d'un spectre unique à chaque objet.

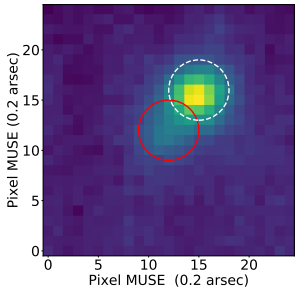
On peut voir sur la figure 2.1 un exemple d'un tel mélange dans les données MUSE. Lorsqu'on estime le spectre de l'objet central (source 1, indiquée en rouge) par intégration sur une ouverture circulaire centrée sur cet objet, on voit apparaître des raies qui appartiennent en fait à plusieurs objets non résolus. En effet, lorsqu'on analyse les images obtenues par bande étroite centrée sur chaque raie, on peut voir que la répartition spatiale varie fortement pour certaines raies. En particulier la raie étudiée dans l'image 2.1d semble clairement provenir d'un autre objet (objet 2 indiqué en pointillé blanc) au vu de sa position spatiale.

Cette analyse peut être confortée par l'étude astrophysique du spectre avec l'apport d'*a priori* physiques permettant d'évaluer la vraisemblance des combinaisons de raie observées. Toutefois, un champ profond comprend plusieurs milliers de sources, avec une proportion non négligeable de sources mélangées, il n'est donc pas envisageable d'explorer de la sorte manuellement chaque situation de mélange pour associer chaque raie à sa véritable source. De plus, les sources émettent également un continuum spectral qui ne peut être démélangé par ces approches simples d'attribution des raies. L'objectif de cette partie est donc de développer une approche non-supervisée permettant de reconstruire pour chaque objet son spectre pur (autrement dit sans contamination de ses voisins).

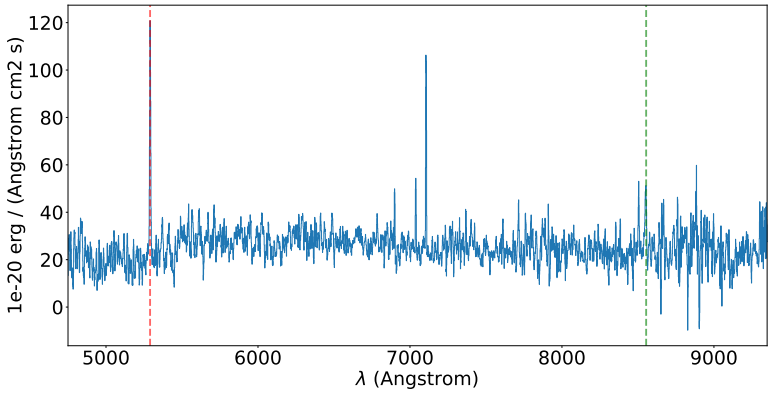
1. ou plus exactement la projection de ces sources sur le ciel observé par MUSE.



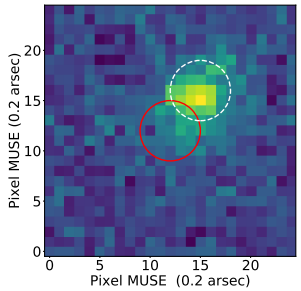
(a) Image issue du télescope Hubble.



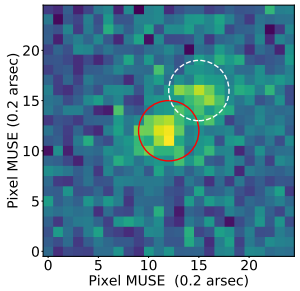
(b) Image blanche MUSE.



(c) Spectre estimé par intégration sur une ouverture circulaire pour la source centrale. On voit apparaître un certain nombre de raies d'émissions.



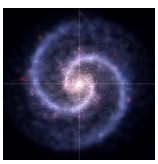
(d) Image "bande étroite" autour de la première raie (en rouge sur le spectre (c)). On voit que cette raie appartient à l'objet 2.



(e) Image "bande étroite" autour de la seconde raie (en vert sur le spectre (c)). On voit que cette raie appartient majoritairement à l'objet central [2].

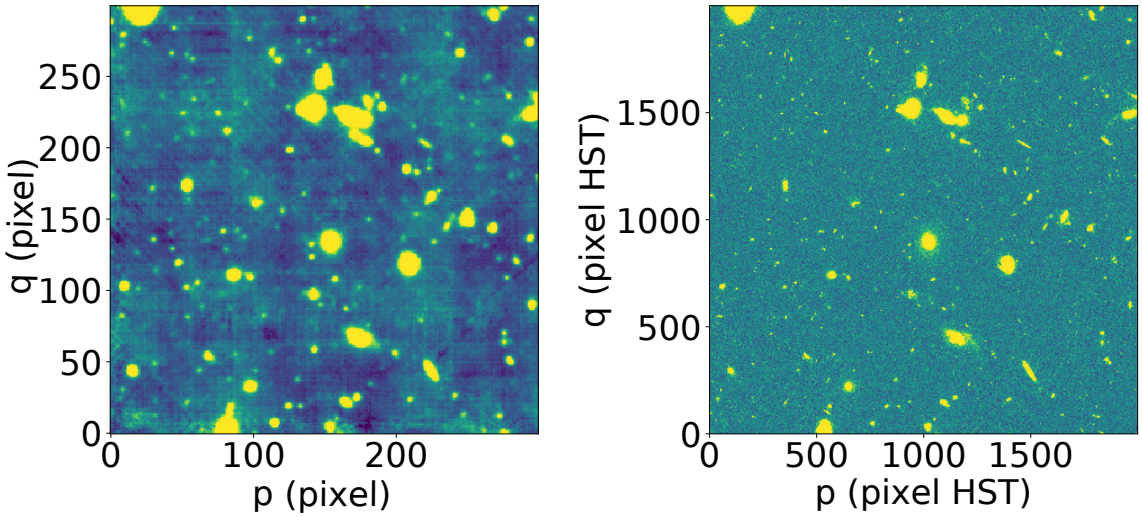
FIGURE 2.1 – Exemple de sources mélangées (source 1 : cercle rouge ; source 2 : cercle pointillé blanc). Par l'analyse d'images de bandes étroites (± 10 feuillets spectraux) autour des raies on voit que le spectre estimé pour la source centrale est notamment contaminé par une raie de la source 2.

2. Le fait que cette raie s'étende spatialement vers l'objet 2 peut s'expliquer de deux façons : soit la raie provient exclusivement d'un composant gazeux du premier objet qui s'étend spatialement, soit le deuxième objet possède également cette raie.

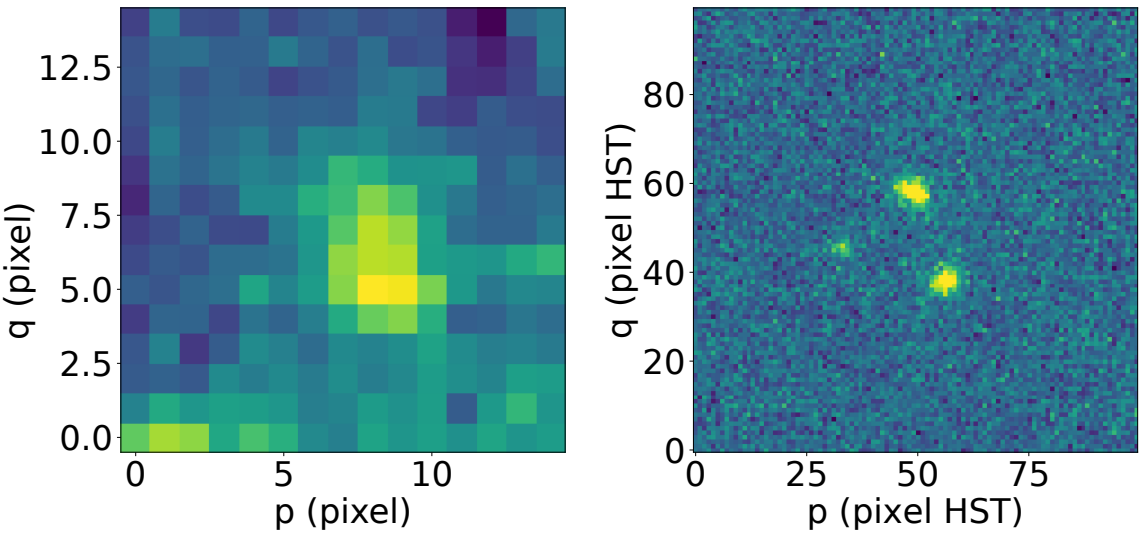


2.1.2 Un atout unique : les observations Hubble

Certains des champs observés par MUSE, notamment les plus profonds tel le *Ultra Deep Field* (UDF), ont également été observés par l'instrument spatial Hubble (*Hubble Space Telescope* ou HST), comme illustré sur l'image 2.2. HST possède non seulement un bien meilleur échantillonnage spatial (un pixel a un côté d'environ 0.03 arcsec) mais également une FSF bien plus piquée (FWHM environ 10 fois plus étroite que celle de MUSE) car non dégradée par l'atmosphère. Toutefois, HST ne possède pas la résolution spectrale de MUSE mais plusieurs filtres passe-bandes qui lui permettent de prendre des images intégrées sur de larges bandes spectrales. En particulier, quatre filtres (désignés par 606W, 775W, 814W et 850LP) couvrent une partie du domaine spectral observée par MUSE, comme illustré par l'image 2.3, et peuvent donc être d'intérêt ici.



(a) Comparaison sur toute l'étendue d'un champ MUSE (1 arcmin^2).



(b) Zoom sur une région présentant des objets mélangés à la résolution MUSE.

FIGURE 2.2 – Comparaison d'un champ MUSE de l'UDF avec une image HST du même champ. L'image blanche de MUSE (somme sur toutes les longueurs d'onde) est à gauche, l'image HST (issue du filtre 775W) à droite.

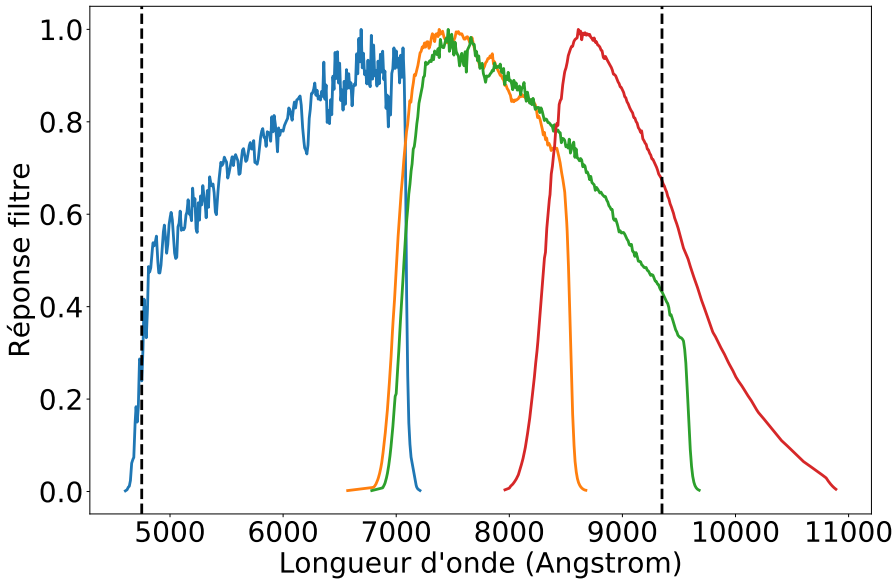


FIGURE 2.3 – Réponse spectrale des différents filtres HST apparaissant dans la bande spectrale de MUSE (indiquée en pointillés noirs). En bleu, le filtre nommé 606W, en orange le filtre 775W, en vert le filtre 814W et en rouge le filtre 850LP.

Nous allons chercher à exploiter les informations spatiales fournies par le HST afin de démélanger les spectres des sources présentes dans un champ MUSE.

2.2 Etat de l'art

2.2.1 Démélange hyperspectral

Au cours de ces dernières années, de nombreuses méthodes ont développées (voir par exemple [KESHAVA et MUSTARD 2002] pour un article de synthèse) pour répondre à la problématique du démélange de données hyperspectrales notamment dans le domaine de la télédétection (observation multi- ou hyperspectrale de la surface planétaire). La plupart de ces méthodes s'appuient sur un modèle de mélange linéaire :

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D} + \boldsymbol{\epsilon}.$$

$\boldsymbol{X} \in \mathbb{R}^{N\times\lambda}$ est un cube hyperspectral (ensemble d'images de taille $p \times q = N$ pixels observées sur λ longueurs d'onde) vectorisé spatialement, chaque ligne de $\boldsymbol{X}$ représentant le spectre d'un pixel. La matrice $\boldsymbol{D} \in \mathbb{R}^{k\times\lambda}$ est une matrice de k composantes spectrales pures, appelées pôles de mélange ou *endmembers*. La matrice $\boldsymbol{U} \in \mathbb{R}^{N\times k}$ contient les abondances de chacune des composantes de la matrice $\boldsymbol{D}$ en chacun des pixels de l'image. Chaque ligne de $\boldsymbol{D}$ contient donc le modèle de spectre d'un pôle et une ligne de $\boldsymbol{U}$ contient les coefficients d'abondance permettant d'exprimer chaque vecteur ligne $\boldsymbol{x}_i$ de $\boldsymbol{X}$ comme une combinaison de lignes (atomes) $\boldsymbol{d}_j$ de $\boldsymbol{D}$:

$$\boldsymbol{x}_i = \sum_j u_{ij}\boldsymbol{d}_j.$$

Notons que dans le cadre classique de la télédétection, la matrice $\boldsymbol{U}$ contient les abondances fractionnaires de chaque pôle au sein d'un pixel, chaque ligne de $\boldsymbol{U}$ doit donc sommer à un (et ses

éléments sont tous positifs). Nous verrons dans le chapitre suivant que dans notre application, la notion d'abondance est peu pertinente, ce qui nous amènera à supprimer cette contrainte de somme à un. Enfin, la matrice $\boldsymbol{\epsilon} \in \mathbb{R}^{n \times \lambda}$ représente les erreurs de reconstruction et le bruit de mesure.

Dans le cadre de problèmes de télédétection, la matrice $\boldsymbol{D}$ peut souvent être formée d'un ensemble (ou dictionnaire) de spectres connus ou de références. Cela ne peut pas s'appliquer pour traiter les données MUSE pour principalement deux raisons :

- Les spectres des galaxies sont a priori inconnus et il n'existe actuellement pas de dictionnaires de spectres de références suffisamment exhaustif.
- La distance des galaxies étant inconnue, la déformation à appliquer à des spectres de référence (du fait du redshift) est également inconnue (voir chapitre 1 pour plus de détails).

Dans notre cas, $\boldsymbol{D}$ doit donc être estimée directement à travers un algorithme d'estimation des "endmembers". On distingue alors deux grandes catégories d'approches pour résoudre ce problème : celle liée au domaine de la séparation aveugle de sources et celle propre au mélange de données hyperspectrales, qui s'appuie sur des arguments géométriques.

La séparation aveugle de sources peut se faire notamment au moyen de l'Independent Component Analysis (ICA) qui a déjà été appliquée à des problèmes de démélange hyperspectral (Bayliss et al. 1998 ; Nascimento et Dias 2005a). L'ICA (Comon 1994) permet de séparer des sources de façon aveugle à l'aide de l'hypothèse d'indépendance statistique des sources. Toutefois les approches type ICA ou IFA (*Independent Factor Analysis*) semblent donner des résultats peu concluants pour ce type de problèmes d'après les études menées dans [Nascimento et Dias 2005a] : dans le cadre de démélange de données hyperspectrales l'ICA ne permet d'obtenir qu'assez approximativement les pôles de mélanges dès que le rapport signal-à-bruit se dégrade. L'explication avancée dans plusieurs de ces études comme [Nascimento et Dias 2005a] ou [Villa et al. 2009], est que le processus d'acquisition des données hyperspectrales ne permet pas de garantir l'indépendance des sources. Par ailleurs, ce type d'approches de séparation aveugle ne permet pas de facilement prendre en compte l'information spatiale donnée par HST. De récents travaux s'intéressent toutefois à l'ajout de contraintes spatiales à l'ICA pour obtenir des approches semi-aveugles, notamment dans le cadre d'applications de fMRI (Lin et al. 2010). D'autres approches de séparation de sources visent à obtenir simultanément les pôles et les abondances en exploitant l'hypothèse de non-négativité des pôles et des abondances. Cette hypothèse permet en effet d'utiliser des techniques de factorisation non-négative de matrices (ou NMF) [Lee et Seung 1999]. Si cette hypothèse de non-négativité est en théorie valable dans le cadre de MUSE (les spectres de lumières des sources étant bien sûr physiquement positifs), les nombreux prétraitements, notamment la soustraction du ciel, rendent en pratique caduque cette hypothèse : l'utilisation de ce type de contrainte aurait pour conséquence de fortement biaiser les solutions en supprimant les composantes négatives dues au bruit, tout en conservant l'influence positive du bruit. L'autre type d'approche classiquement utilisée en télédétection consiste à remarquer que les observations d'une scène se trouvent à l'intérieur d'un simplexe dont les sommets correspondent aux pôles de mélange. Toutefois, ces méthodes géométriques d'estimation des pôles d'abondances tel que VCA (*Vertex Component Analysis*) [Nascimento et Dias 2005b] ou NFINDR [Winter 1999] reposent le plus souvent sur une hypothèse de présence dans les observations de "pixels purs", c'est à dire de pixels contenant du signal provenant d'une seule source. Cette hypothèse est difficile à garantir ici pour toutes les sources. De plus ces méthodes ne sont pas adaptées en terme de nombre de

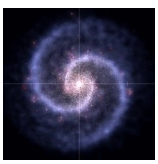
pôles de mélange, bien plus élevé dans un champ MUSE (plusieurs centaines contre quelques dizaines en télédétection) et en terme de dimension spectrale (quelques centaines de longueurs d'onde en télédétection contre plusieurs milliers dans notre application).

Ainsi, bien que la problématique étudiée soit très proche du démélange linéaire étudié classiquement en télédétection, des différences conséquentes apparaissent du fait de l'application souhaitée. En effet dans le cadre des données astrophysiques MUSE, le nombre d'objets (pôles) considérés est un ou deux ordres de grandeur plus grand que dans la plupart des applications de télédétection. Les rapports signaux-à-bruit sont également très différents, les données issues de la télédétection classique étant relativement très peu bruitées comparativement aux données astrophysiques. Enfin l'objectif du démélange est également de nature différente : la plupart des applications de démélange en télédétection s'intéressent en premier lieu aux cartes d'abondance et non aux spectres des pôles de mélanges (souvent connus) en tant que tels. Au contraire ici, nous sommes avant tout intéresser par l'analyse chimique des objets et on souhaite donc en premier lieu estimer les spectres.

Notons par ailleurs qu'une autre particularité du travail présenté ici est qu'il possède une contrainte applicative forte. L'objectif est en effet de fournir un algorithme suffisamment robuste et rapide pour en permettre une utilisation régulière par les astronomes du CRAL. Il s'agit donc de privilégier une solution rapide et d'éviter une modélisation trop complexe et coûteuse en temps de calcul, les jeux de données à traiter étant de taille importante (de l'ordre de 3Go et 300 millions de voxels par observation). On cherche également à limiter le plus possible le nombre de paramètres à régler pour tendre vers une résolution non-supervisée.

2.2.2 Pansharpening

A partir des données Hubble, une autre façon d'appréhender le problème de limitation spatiale de MUSE est d'utiliser des approches de type super-résolution spatiale ou *pansharpening*. Le *pansharpening* est en effet un champ d'étude très exploré pour les données hyperspectrales, qui cherche à répondre à la problématique suivante : à partir de données à haute résolution spatiale mais faible résolution spectrale (image panchromatique ou données multispectrales) et de données à faible résolution spatiales mais haute résolution spectrale (données hyperspectrales), comment reconstruire un jeu de données à haute résolution spatiale et spectrale ? Comme exposé dans le papier de synthèse [Loncan et al. 2015], de nombreuses méthodes ont été développées ces dernières années pour répondre à cette problématique de pansharpening, notamment dans le domaine de la télédétection. D'après [Loncan et al. 2015], une des méthodes les plus performantes (Wei et al. 2015) s'appuie sur une régularisation (sous la forme d'un *a priori* Bayesien) par contrainte de parcimonie spatiale. Les images naturelles sont en effet souvent parcimonieuses sur un dictionnaire adapté. D'autres approches s'appuient sur des hypothèses de non-négativité comme la méthode *Coupled Nonnegative Matrix Factorization* (ou CNMF) [Yokoya et al. 2012]. Toutefois, dans le cadre de l'application MUSE, la construction d'un cube super-résolu (résultat d'une approche pansharpening) n'est pas désirée, d'une part à cause de la taille des jeux de données (un cube MUSE ramené à la résolution spatiale du HST ferait de l'ordre de 150 Go et 16 milliards de voxels) et d'autre part car le principal objectif scientifique est l'estimation des spectres des sources. Nous nous sommes donc concentrés sur une méthode de démélange linéaire exploitant les informations HST sans construction d'un cube super-résolu. Des approches permettant d'obtenir un cube super-résolu ont toutefois été explorées et sont présentées brièvement dans les perspectives.



Remarque 1
Dans le cadre de MUSE, on entendra par pôles de mélange *les sources suffisamment brillantes (galaxies, étoiles) et non le signal contenu notamment dans les structures extragalactiques, plus diffus et difficilement modélisable par un spectre unique. Ce type de signal diffus est donc négligé dans un premier temps. Il est à noter qu'une forte approximation est faite quand à la variabilité spectrale des sources brillantes étendues (galaxies proches) : du fait de la dynamique de la source le spectre peut être plus ou moins dilaté spectralement d'un pixel à l'autre par effet Doppler. Contrairement aux variabilités classiquement étudiées en télédétection (pour prendre en compte les changements d'ensoleillement par exemple), cette dilatation spectrale ne peut pas aisément être intégrée dans le modèle linéaire considéré. Elle n'est donc pas prise en compte pour l'instant, i.e. les sources sont caractérisées par un spectre unique.*

Résumé

On s'appuie sur un modèle de mélange linéaire :
- Des signatures spectrales/endmembers
- Des cartes d'intensité/abondances

Dans l'état de l'art :
- La plupart des méthodes de démélange cherchent principalement à estimer les cartes d'abondances
- Les méthodes d'extraction d'*endmembers* ne sont pas adaptées à de telles dimensions (3600 longueurs d'onde), ni assez robustes aux niveaux de bruit considérés.
- Les méthodes de séparation de sources (NMF, ICA) ne sont pas adaptées ici.

L'application étudiée permet l'utilisation d'une source d'informations spatiales forte : les images HST. On va donc chercher à exploiter ces informations dans notre approche, qui vise à être robuste, non-supervisée et peu coûteuse d'un point de vue calculatoire.

Chapitre 3

Méthode de démélange proposée

Sommaire

Ce chapitre expose la méthode de démélange proposée pour répondre aux spécificités de MUSE. Après l'introduction des notations nécessaires dans la section 3.1, une première méthode est présentée dans la section 3.2. Cette première méthode montre rapidement ses limites lorsque le problème de mélange devient mal conditionné. Pour surmonter ces limites une régularisation est proposée dans la section 3.3 et validée sur données simulées.

3.1 Notations

On note par une majuscule en gras ($\boldsymbol{M}$) une matrice composée des éléments $m_{i,j}$ et par une minuscule en gras un vecteur ($\boldsymbol{v}$). La notation diag(s_j) désigne une matrice diagonale dont les éléments diagonaux sont les s_j. La notation $||\boldsymbol{M}||_2$ indique la norme de Frobenius de la matrice $\boldsymbol{M}$ soit $||\boldsymbol{M}||_2 = \sqrt{\sum_{i,j} m_{i,j}^2}$. La trace d'une matrice $\boldsymbol{M}$ est notée Tr($\boldsymbol{M}$). La pseudo inverse d'une matrice $\boldsymbol{M}$ est notée $\boldsymbol{M}^+$. Le produit de Hadamard est défini entre une matrice $\boldsymbol{A}$ et un vecteur $\boldsymbol{b}$ par

$$\boldsymbol{A} \circ \boldsymbol{b} = \begin{pmatrix} a_{11} \cdot b_1 & \cdots & a_{1n} \cdot b_1 \\ \vdots & \ddots & \vdots \\ a_{m1} \cdot b_m & \cdots & a_{mn} \cdot b_m \end{pmatrix}.$$

On note $\boldsymbol{Y}$ la matrice de données issue de MUSE de dimension $n \times \lambda$ où $n = pq$ est le nombre de pixels (≈ 90000) et λ est le nombre de bande spectrales ou feuillets (≈ 3600). On note$\boldsymbol{Z}$ la matrice de données issue de HST de dimension $N \times \Lambda$ avec $\Lambda(\approx 4) \ll \lambda$ et $n \ll N(\approx 6 \times 10^5)$.

On note $\boldsymbol{X}$ le champ de données (inconnu) super-résolu spatialement et spectralement de dimension $N \times \lambda$.

On considère que les données observées par MUSE correspondent au champ de données super-résolu $\boldsymbol{X}$ ayant subi une dégradation spatiale et l'addition d'un bruit d'observation et de mesure $\mathcal{N}$. La dégradation spatiale est la multiplication d'une opération de convolution (par la FSF) et d'une opération de sous-échantillonnage. La FSF variant avec la longueur d'onde, on a le modèle suivant :

$$\text{Pour } 1 \leq l \leq \lambda, \quad \underset{n\times 1}{\boldsymbol{Y}_l} = \underset{n\times N}{\boldsymbol{B}_l} \times \underset{N\times 1}{\boldsymbol{X}_l} + \underset{n\times 1}{\mathcal{N}} \tag{3.1}$$

avec $\boldsymbol{B}_l$ la matrice de dégradation spatiale (sous-échantillonnage et convolution) et se décompose sous la forme $\boldsymbol{B}_l = \boldsymbol{S} \times \boldsymbol{C}_l$ où $\boldsymbol{C}_l$ est la matrice de convolution par la FSF à la longueur d'onde l et $\boldsymbol{S}$ la matrice de sous-échantillonnage.

On considère l'observation HST, composée des images issues des différents filtres, comme le résultat d'une dégradation spectrale de $\boldsymbol{X}$:

$$\underset{N\times\Lambda}{\boldsymbol{Z}} = \underset{N\times\lambda}{\boldsymbol{X}} \times \underset{\lambda\times\Lambda}{\boldsymbol{A}} \tag{3.2}$$

avec $\boldsymbol{A}$ la matrice des réponses des différents filtres HST.

On note également $\boldsymbol{H} \in \mathbb{R}^{N\times k}$ la matrice de segmentation à la résolution HST. Cette matrice de segmentation est obtenue ici à partir des méthodes proposées dans [Rafelski et al. 2015]. Les éléments de $\boldsymbol{H}$ sont définis par :

$$h_{ij} = \begin{cases} 1 \text{ si la source } j \text{ est présente sur le pixel } i, \\ 0 \text{ sinon.} \end{cases}$$

Remarque 2
Notons qu'on s'affranchit ici des problématiques d'alignement spatial, les données MUSE et HST ayant déjà été spatialement positionnées avec précision sur une grille de coordonnées célestes.

Remarque 3
On peut également noter qu'on ne prend pas en compte ici l'influence de la LSF qui est négligeable par rapport aux structures d'intérêt (raies d'émission ou d'absorption) des spectres pour l'analyse astrophysique, la largeur de la LSF étant un ordre de grandeur plus petite que la plupart des raies. De manière similaire, on peut considérer les données HST comme quasi-idéales spatialement (en pratique la FSF du HST étant connue, nous la prendrons en compte lors de l'opération de dégradation spatiale de HST à la résolution MUSE).

3.2 Approche directe : inversion par moindres carrés ordinaires

3.2.1 Hypothèses/Modélisation

On fait ici les hypothèses suivantes.

A1 (Séparation spatiale)**.** *Toutes les sources présentes sur les images HST peuvent être séparées spatialement.*

A2 (Invariabilité spatiale)**.** *La répartition spatiale d'une source varie peu en fonction de la longueur d'onde.*

A3 (Invariabilité spectrale)**.** *Les sources sont modélisables par un unique spectre.*

A4 (Invariabilité de la FSF)**.** *La fonction de transfert HST-MUSE est connue, est invariable spatialement, et peut être supposée constante sur une certaine largeur de bande spectrale.*

L'hypothèse A1 est valide pour la très grande majorité des sources présentes sur les images HST. Il existe en pratique un certain nombre de cas où les sources ne sont pas séparables spatialement sur HST, sources pour lesquels la méthode de démélange proposée ici ne sera donc pas fonctionnelle.

Les hypothèses A2 et A3 sont également des approximations mais permettent la séparabilité des sources. Certaines sources, notamment les galaxies proches et résolues spatialement sur MUSE possèdent des champ de vitesses qui vont déformer les spectres observés, brisant l'hypothèse A3. D'autres sources possèdent des émissions à certaines longueurs d'onde différant fortement du continuum, voire ne possèdent pas de continuum, rendant l'information de HST peu pertinente. L'étude de ces sources spatialement étendues sur quelques longueurs d'onde est abordée dans la partie II de ce manuscrit.

L'hypothèse A4 est en pratique bien vérifiée. La FSF du HST est désormais bien connue et la FSF de MUSE a fait l'objet de plusieurs travaux au sein du consortium (Villeneuve et al. 2011 ; Carfantan 2014). Comme décrit dans le chapitre 1, elle est modélisée par une fonction Moffat 2D circulaire (Moffat 1969) d'expression

$$F_\lambda(r) = \frac{\beta_\lambda - 1}{\pi\alpha^2}\left(1 + \frac{r^2}{\alpha_\lambda^2}\right)^{-\beta_\lambda},$$

r étant la distance au centre, où le paramètre d'échelle α_λ varie lentement en fonction de la longueur d'onde et le paramètre de forme β_λ. La variation du paramètre d'échelle α_λ est suffisamment lente (voir figure 1.6) pour pouvoir considérer la FSF constante sur plusieurs dizaines voir centaines de bandes spectrales MUSE.

L'hypothèse A2 permet notamment de se placer dans le cadre d'un modèle de mélange linéaire :

$$\boldsymbol{X} \approx \underset{N\times k}{\boldsymbol{U}} \times \underset{k\times\lambda}{\boldsymbol{D}}$$

où k est le nombre de sources présentes dans les données ($k(\sim 1000) \ll N$), $\boldsymbol{D} \in \mathbb{R}^{k\times\lambda}$ est la matrice des spectres des sources, et $\boldsymbol{U} \in \mathbb{R}^{n\times k}$ la matrice d'abondance ou plus exactement d'intensité. Notons en effet que dans notre application le terme d'abondance est un abus de langage. Il s'agit en réalité des intensités de chaque source. En effet, contrairement aux applications de télédétection classiques nous nous plaçons dans un cadre où on recherche un ensemble d'objets qui se détachent d'un fond et qui possèdent de fortes variations d'intensité. Nous ne nous situons en particulier plus dans le cadre de la contrainte de somme à un des lignes de $\boldsymbol{U}$, classiquement exploitée en télédétection.

Toutes ces hypothèses vont nous permettre de déduire une matrice d'intensité $\boldsymbol{U}$ à l'aide des images HST. A partir des données MUSE et de cette matrice $\boldsymbol{U}$, on va donc chercher à estimer directement la matrice des spectres $\boldsymbol{D}$, sans reconstruire explicitement $\boldsymbol{X}$.

La démarche est alors la suivante :

1. On travaille indépendamment sur chaque filtre i du HST (par simplicité d'implémentation).

2. Pour chaque image (vectorisée) $\underbrace{\boldsymbol{Z}_i}_{N\times 1}$ du HST :
   (a) On obtient la matrice d'intensité $\boldsymbol{U}_i$ à partir de $\boldsymbol{Z}_i$.
   (b) Pour chaque longueur d'onde l, on dégrade spatialement $\boldsymbol{U}_i$ pour le ramener à la résolution spatiale de MUSE.
   (c) Pour chaque longueur d'onde l, on inverse le système $\underbrace{\boldsymbol{Y}_l}_{n\times 1} = \underbrace{\widetilde{\boldsymbol{U}}_{i,l}}_{n\times k}\underbrace{\boldsymbol{D}_{i,l}}_{k\times 1}$ ce qui revient à chercher $\widehat{\boldsymbol{D}}_{i,l}$ solution de

$$\widehat{\boldsymbol{D}}_{i,l} = \underset{\boldsymbol{D}_{i,l}}{\operatorname{argmin}} ||\boldsymbol{Y}_l - \widetilde{\boldsymbol{U}}_{i,l}\boldsymbol{D}_{i,l}||_2^2. \tag{3.3}$$

3. On combine ensuite les différentes estimations faites sur chaque bande HST en faisant une moyenne pondérée par les réponses des filtres :

$$\widehat{\boldsymbol{D}}_l = \sum_i a_{i,l}\widehat{\boldsymbol{D}}_{i,l}, \quad 1 \leq l \leq \lambda, \tag{3.4}$$

   où $a_{i,l}$ correspond à la réponse spectrale du filtre i à la longueur d'onde l.

3.2.2 Construction de la matrice d'intensité

La première étape (a) se fait à l'aide de la carte de segmentation des sources HST (issue de Rafelski et al. 2015). On construit à partir de cette carte une matrice $\underbrace{\boldsymbol{H}}_{N\times k}$, chaque colonne j de $\boldsymbol{H}$ correspond au masque binaire de présence/absence de la source j dans le champ HST vectorisé. On effectue ensuite un produit de Hadamard entre $\boldsymbol{H}$ et l'image HST $\boldsymbol{Z}_i$ pour obtenir la matrice d'intensité $\boldsymbol{U}_i$:

$$\underbrace{\boldsymbol{U}_i}_{N\times k} = \underbrace{\boldsymbol{H}}_{N\times k} \circ \underbrace{\boldsymbol{Z}_i}_{N\times 1} \tag{3.5}$$

Lors de la deuxième étape (b), on commence par effectuer une multiplication par la matrice de convolution entre HST et MUSE. Le noyau de convolution de HST à MUSE K_{HM} est obtenue à partir des noyaux de convolution du HST K_H et de MUSE K_M, supposés connus. Puisque par définition $K_M = K_H * K_{HM}$, dans le domaine de Fourier on a $\mathcal{F}(K_M) = \mathcal{F}(K_H)\mathcal{F}(K_{HM})$, avec $\mathcal{F}(.)$ la transformée de Fourier. On obtient donc le noyau de convolution de HST à MUSE par

$$K_{HM} = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(K_M)}{\mathcal{F}(K_H)}\right).$$

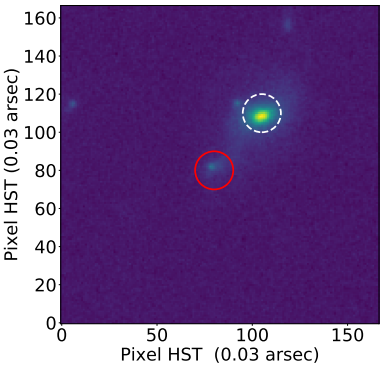
Puis on ré-échantillonne le résultat de la convolution à la résolution de MUSE, à l'aide d'une interpolation linéaire (rappelons que l'alignement de l'image HST et du champ MUSE est assuré en amont par l'utilisation d'un système de coordonnées célestes commun).

On a donc l'expression suivante :

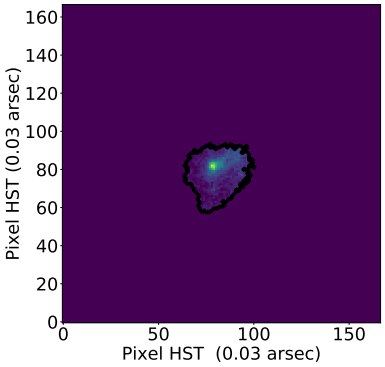
$$\underbrace{\widetilde{\boldsymbol{U}}_{i,l}}_{n\times k} = \underbrace{\boldsymbol{B}_l}_{n\times N}\underbrace{\boldsymbol{U}_i}_{N\times k}, \tag{3.6}$$

où $\widetilde{\boldsymbol{U}}_{i,l}$ est la matrice d'intensité à la résolution MUSE.

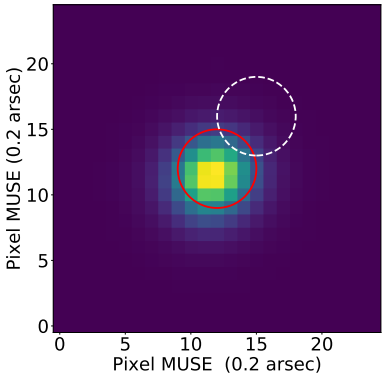
La figure 3.1 illustre cette construction des cartes d'intensité. Pour chaque objet défini par la carte de segmentation, on peut obtenir sa carte d'intensité propre, même si les objets ne sont pas séparables sur les données MUSE. Notons que la carte de segmentation (qui cherche à indiquer l'extension maximale de chaque galaxie) est identique pour tous les filtres HST. En revanche le profil d'intensité de chaque galaxie varie légèrement d'un filtre à l'autre (marquant le fait que les différents éléments chimiques au sein d'une galaxie ne sont pas identiquement répartis spatialement).



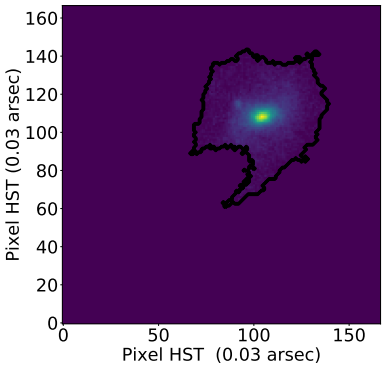
(a) Image HST



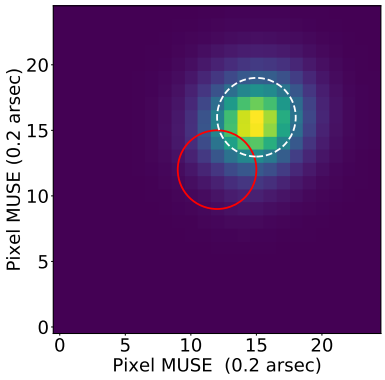
(b) Carte d'intensité de l'objet 1 à la résolution HST



(c) Carte d'intensité de l'objet 1 à la résolution MUSE



(d) Carte d'intensité de l'objet 2 à la résolution HST



(e) Carte d'intensité de l'objet 2 à la résolution MUSE

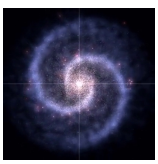
FIGURE 3.1 – Les cartes d'intensité sont obtenues à partir de l'image HST et de la carte de segmentation, et sont ensuite dégradées à la résolution MUSE

3.2.3 Estimation des spectres

La troisième étape (c) se fait alors par résolution au sens des moindres carrés du critère (3.3), avec k inconnues pour n équations, $k \ll n$. Le problème est sur-déterminé et a donc une unique solution dont l'expression analytique est la suivante :

$$\widehat{\boldsymbol{D}}_{i,l} = (\widetilde{\boldsymbol{U}}_{i,l}^T \widetilde{\boldsymbol{U}}_{i,l})^{-1} \widetilde{\boldsymbol{U}}_{i,l}^T \boldsymbol{Y}_l. \tag{3.7}$$

Nous verrons par la suite dans la section 3.3 qu'une régularisation s'avère toutefois nécessaire lorsque les sources sont fortement mélangées.



Les deuxième et troisième étapes sont coûteuses à exécuter sur les 3600 longueurs d'onde de MUSE. En pratique, la FSF variant lentement avec la longueur d'onde, on peut opérer par bloc (de typiquement 200 ou 300 bandes spectrales), en estimant une FSF moyenne sur l'intervalle considéré. Cela permet de réduire considérablement le coût calculatoire puisqu'il suffit de pseudo-inverser une seule matrice $\widetilde{\boldsymbol{U}}_{i,l}$ par bloc l. Notons qu'on obtient toutefois bien une estimation du spectre pour chaque longueur d'onde, seule la matrice d'intensité est supposée constante sur le bloc spectral considéré (hyp. A2).

Par ailleurs, avec chaque spectre estimé est donnée la variance de l'estimation. On propage pour cela lors de l'étape finale les informations de variance fournie avec la matrice de données MUSE. Comme décrit dans le chapitre 1, les données MUSE sont fournies avec une estimation de la variance du bruit, estimée à partir des capteurs et propagée le long de la chaîne de traitement. Bien que les données soient corrélées (notamment du fait de divers ré-échantillonnages avec interpolation), les informations de covariance ne sont pas propagées depuis les capteurs (principalement pour des raisons de taille) et leur estimation à partir des données finales est peu aisée. On assimile donc la matrice de variance-covariance des données (dont les termes diagonaux sont inconnus) à une matrice de variance diagonale $\underbrace{\boldsymbol{\Sigma}_l}_{n\times n}$ pour chaque feuillet MUSE $\boldsymbol{Y}_l$. Puisque $\widehat{\boldsymbol{D}}_{i,l} = (\widetilde{\boldsymbol{U}}_{i,l}^T\widetilde{\boldsymbol{U}}_{i,l})^{-1}\widetilde{\boldsymbol{U}}_{i,l}^T\boldsymbol{Y}_l$ avec $\mathrm{var}(\boldsymbol{Y}_l) = \boldsymbol{\Sigma}_l$ la variance de l'estimateur $\widehat{\boldsymbol{D}}_{i,l}$ est

$$\underbrace{\mathrm{var}(\widehat{\boldsymbol{D}}_{i,l})}_{k\times k} = (\widetilde{\boldsymbol{U}}_{i,l}^T\widetilde{\boldsymbol{U}}_{i,l})^{-1}\widetilde{\boldsymbol{U}}_{i,l}^T\boldsymbol{\Sigma}_l\widetilde{\boldsymbol{U}}_{i,l}(\widetilde{\boldsymbol{U}}_{i,l}^T\widetilde{\boldsymbol{U}}_{i,l})^{-1}. \tag{3.8}$$

Cette variance peut ainsi être prise en compte lors de l'analyse astrophysique de ces spectres.

La figure 3.2 illustre la résolution du problème inverse : à partir des données MUSE et des cartes d'intensité issues du HST, on cherche estimer la contribution de chaque objet à chaque longueur d'onde. La méthode complète est résumée par l'algorithme 1.

Algorithme 1 Procédure d'estimation des spectres par moindres carrés

Entrée : Données MUSE $\boldsymbol{Y}$, données HST $\boldsymbol{Z}$ et carte de segmentation $\boldsymbol{H}$

for filtre HST i **do**
  Calcul de $\boldsymbol{U}_i$ à l'aide de l'équation (3.5) ▷ Construction de la matrice d'intensité
  for bande spectrale l **do**
    Calcul de $\widetilde{\boldsymbol{U}}_{i,l}$ à l'aide de (3.6) ▷ Transformation à la résolution de MUSE
    Calcul de $\widehat{\boldsymbol{D}}_{i,l}$ à l'aide de l'équation (3.7) ▷ Estimation des spectres
    Calcul de $\mathrm{var}(\widehat{\boldsymbol{D}}_{i,l})$ à l'aide de l'équation (3.8) ▷ Variances des estimations
  Construction de $\widehat{\boldsymbol{D}}_i$, $\mathrm{var}(\widehat{\boldsymbol{D}}_i)$ ▷ Concaténation spectrale
Calcul de $\widehat{\boldsymbol{D}}$, $\mathrm{var}(\widehat{\boldsymbol{D}})$ à l'aide de (3.4) ▷ Combinaison des filtres HST
Sortie : $\widehat{\boldsymbol{D}}$, $\mathrm{var}(\widehat{\boldsymbol{D}})$ ▷ Spectres des sources et variances

3.2.4 Validation sur données simulées

3.2.4.1 Construction des données

Afin de valider le bon fonctionnement de la méthode proposée, on construit une série de données avec des sources de plus en plus proches spatialement. On construit pour cela un cube

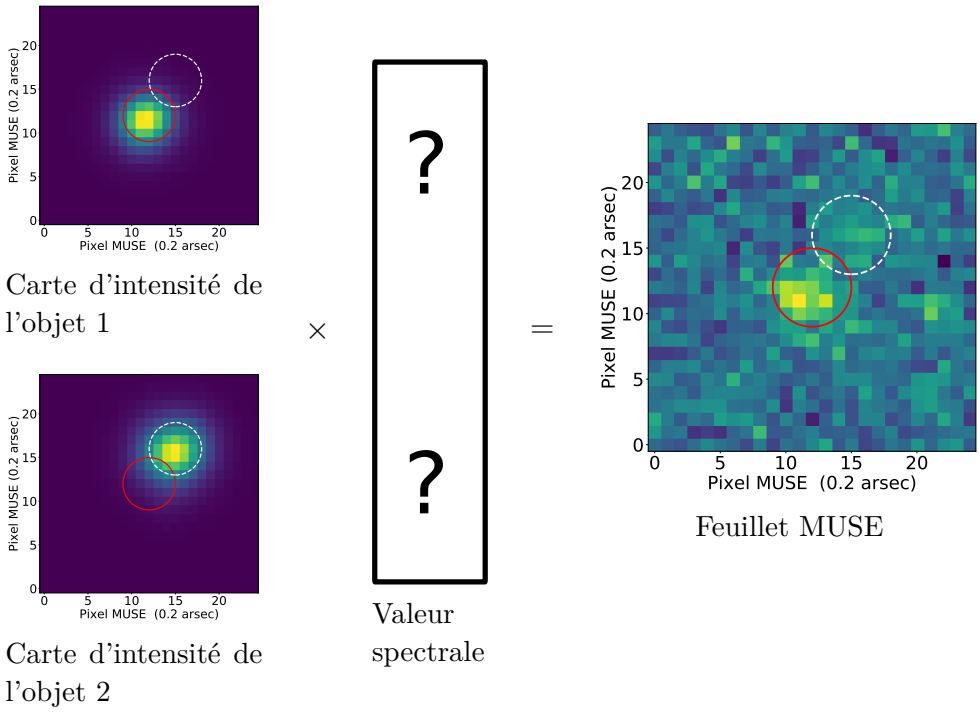


FIGURE 3.2 – A partir des données MUSE et des cartes d'intensité issues du HST, on cherche à estimer l'intensité spectrale de chaque objet à une longueur d'onde donnée.

de données (de taille 160×160) à haute résolution (qu'on appellera données HST par abus de langage) qui est ensuite dégradé en suivant le modèle de MUSE (même sous-échantillonnage et noyau de convolution) pour obtenir des données à basse résolution spatiale (qu'on appellera données MUSE par abus de langage). Un bruit additif gaussien est ensuite appliqué, légèrement corrélé spatialement et spectralement par l'application d'un noyau spatial de taille 3 par 3, ici $\begin{pmatrix} 0 & 0.1 & 0 \\ 0.1 & 1 & 0.1 \\ 0 & 0.1 & 0 \end{pmatrix}$ et un noyau spectral de taille 3, ici $\begin{pmatrix} 0.1 & 1 & 0.1 \end{pmatrix}$.

On peut voir sur l'image 3.3 l'image blanche des cubes simulés à la résolution MUSE et HST en fonction de l'écartement des sources.

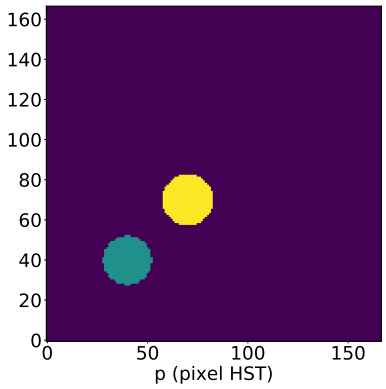
Les spectres des objets sont construits à partir de spectres issus de données réelles MUSE et sont illustrés en figure 3.4. On peut voir une raie caractéristique pour chaque objet ainsi qu'un continuum variant plus ou moins rapidement le long de chaque spectre.

3.2.4.2 Indicateurs

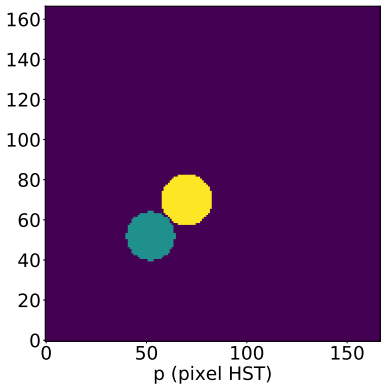
La difficulté de la tâche de démélange est mesurée à l'aide du nombre de conditionnement c de la matrice d'intensité $\widehat{\boldsymbol{U}}$ à la résolution spatiale MUSE :

$$c = \frac{s_{max}(\widehat{\boldsymbol{U}})}{s_{min}(\widehat{\boldsymbol{U}})}, \tag{3.9}$$

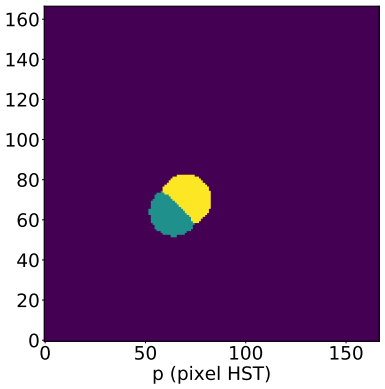
où $s_{max}(\widehat{\boldsymbol{U}})$ et $s_{min}(\widehat{\boldsymbol{U}})$ sont les valeurs singulières maximale et minimale de la matrice d'intensité. Le nombre de conditionnement c est une borne de l'erreur d'estimation de $\widehat{\boldsymbol{D}}$ relativement à une perturbation sur les données $\boldsymbol{Y}$. Ainsi plus c est proche de 1, mieux le problème est



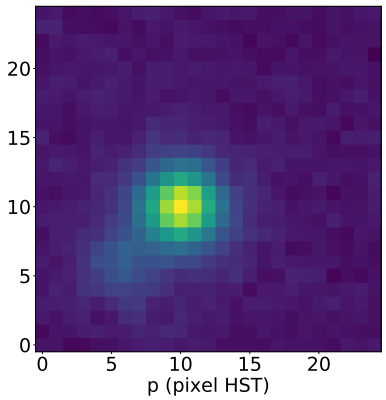
(a) Carte de segmentation HST pour des sources très éloignés ($c = 1.2$)



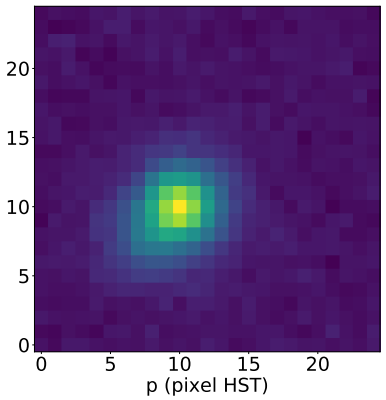
(b) Carte de segmentation HST pour des sources assez proches ($c = 2.5$)



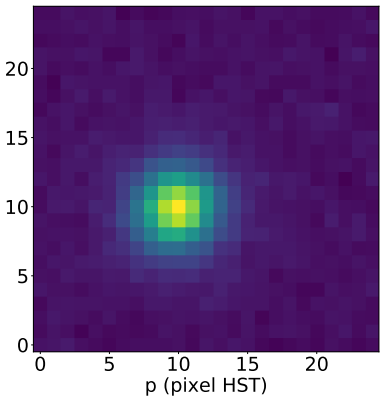
(c) Carte de segmentation HST pour des sources très proches ($c = 4$)



(d) Image blanche MUSE pour des sources très éloignés ($c = 1.2$)



(e) Image blanche MUSE pour des sources assez proches ($c = 2.5$)



(f) Image blanche MUSE pour des sources très proches ($c = 4$)

FIGURE 3.3 – Cartes de segmentation "HST" (lignes du haut) et images blanches "MUSE" (ligne du bas) en fonction de l'écartement des sources (colonnes). Le nombre de conditionnement c est défini dans l'eq. (3.9)

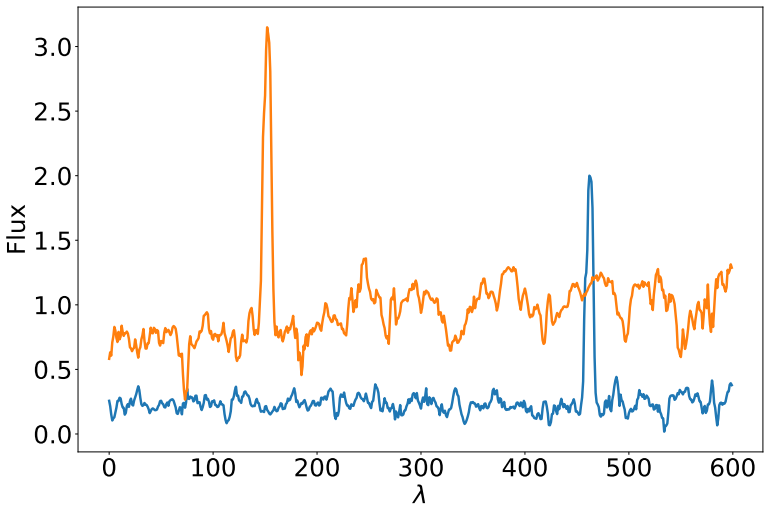


FIGURE 3.4 – Spectres non bruités construits à partir de spectres MUSE pour les deux sources.

conditionné et l'inversion sera stable. Au contraire des fortes valeurs de c indiquent que la matrice à inverser est mal conditionnée et le système devient instable (toute perturbation de $\boldsymbol{Y}$ entraînera de fortes variations de $\widehat{\boldsymbol{D}}$ qui sur-ajustera le bruit). On peut voir sur la figure 3.5 que c est directement lié à l'éloignement des sources. En effet, la valeur de c dépend de la colinéarité entre les colonnes de $\widetilde{\boldsymbol{U}}$ i.e. du recouvrement spatial entre les sources.

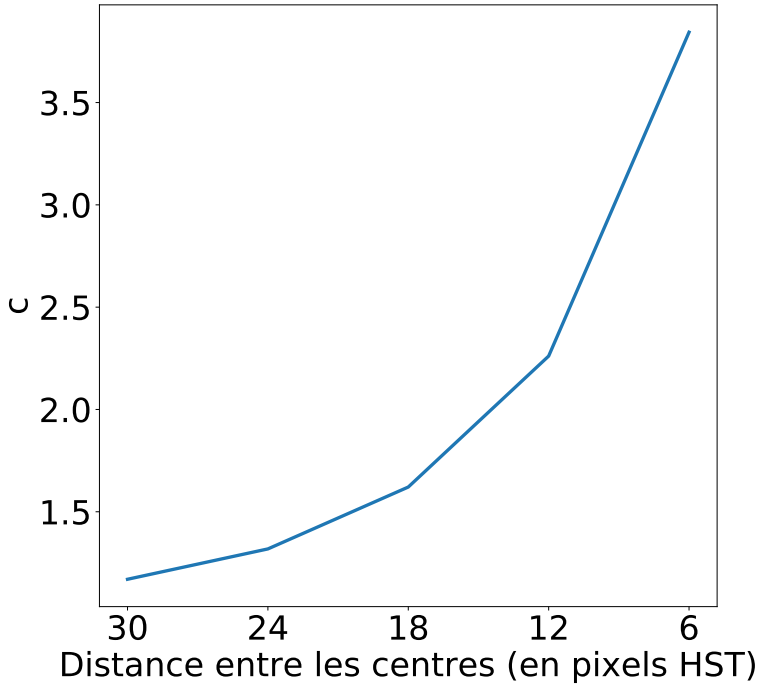
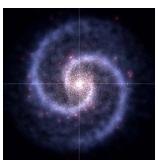


FIGURE 3.5 – Evolution du conditionnement du problème (3.3) lorsque la distance entre les centres des sources diminue.

On mesure ensuite sur les spectres estimés leur variance moyenne

$$v = \frac{1}{k}\sum_{j=1}^{k} \mathrm{var}(\widetilde{\boldsymbol{d}}_j), \tag{3.10}$$



où $\widetilde{\boldsymbol{d}}_j$ est le spectre j privé de ses raies, la fidélité à la vérité terrain (moyenne des corrélations entre chaque spectre estimé $\widehat{\boldsymbol{d}_j}$ et son spectre de référence $\boldsymbol{d}_j$)

$$f = \frac{1}{k}\sum_{j=1}^{k} \frac{\langle \widehat{\boldsymbol{d}_j}, \boldsymbol{d}_j \rangle}{||\widehat{\boldsymbol{d}_j}||.||\boldsymbol{d}_j||}, \tag{3.11}$$

et l'intercorrélation

$$ic = \frac{\langle \widehat{\boldsymbol{d}_0}, \widehat{\boldsymbol{d}_1} \rangle}{||\widehat{\boldsymbol{d}_0}||.||\widehat{\boldsymbol{d}_1}||} \tag{3.12}$$

entre les 2 spectres estimés.

3.2.4.3 Résultats

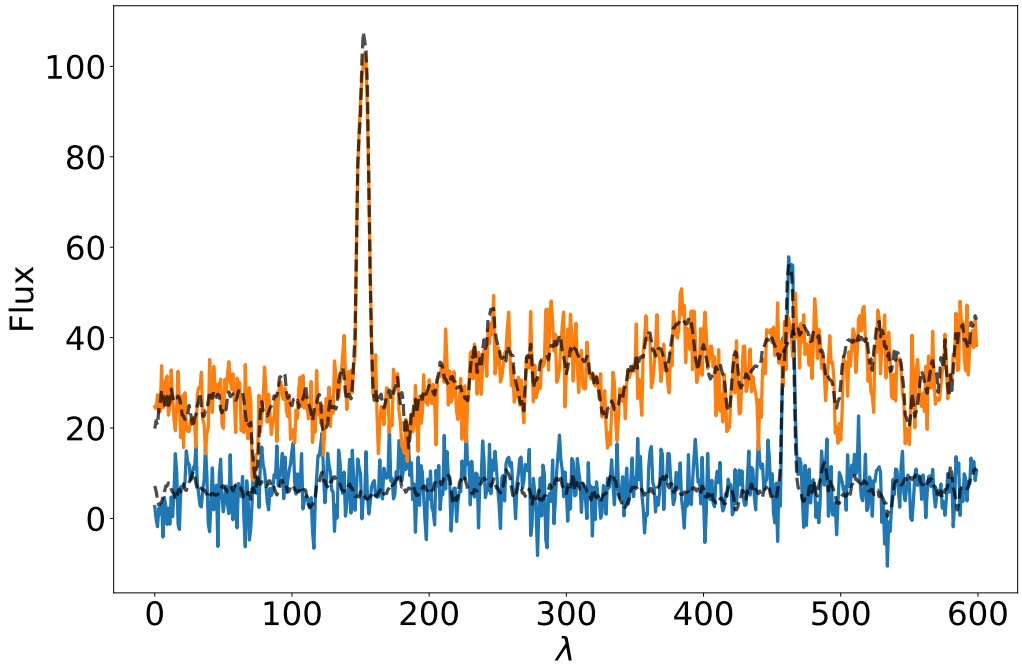
Regardons dans un premier temps les résultats lorsque la tâche de démélange est relativement bien conditionnée ($c \approx 2.5$).

Les spectres estimés sont comparés aux spectres de référence sur la figure 3.6. On peut voir en 3.6(a) que la méthode de démélange permet bien d'obtenir les spectres de chacune des deux sources. Notons qu'une approche plus directe, classiquement utilisée en astronomie et consistant simplement à sommer les spectres pixeliques sur une ouverture circulaire correspondant au support spatial de la source, ne permet pas de démélanger correctement les spectres : on peut voir sur la figure 3.6(b) pour chaque spectre la contamination par les raies du spectre voisin. La variance des résidus est affichée sur la figure 3.7. L'erreur de reconstruction est ici comparable aux fluctuations du bruit, dans ce cas simulé suivant parfaitement le modèle de mélange linéaire. Le démélange permet également de débruiter en partie les spectres comme illustré sur la figure 3.8 où les spectres estimés au centre de la source 2 sont comparés.

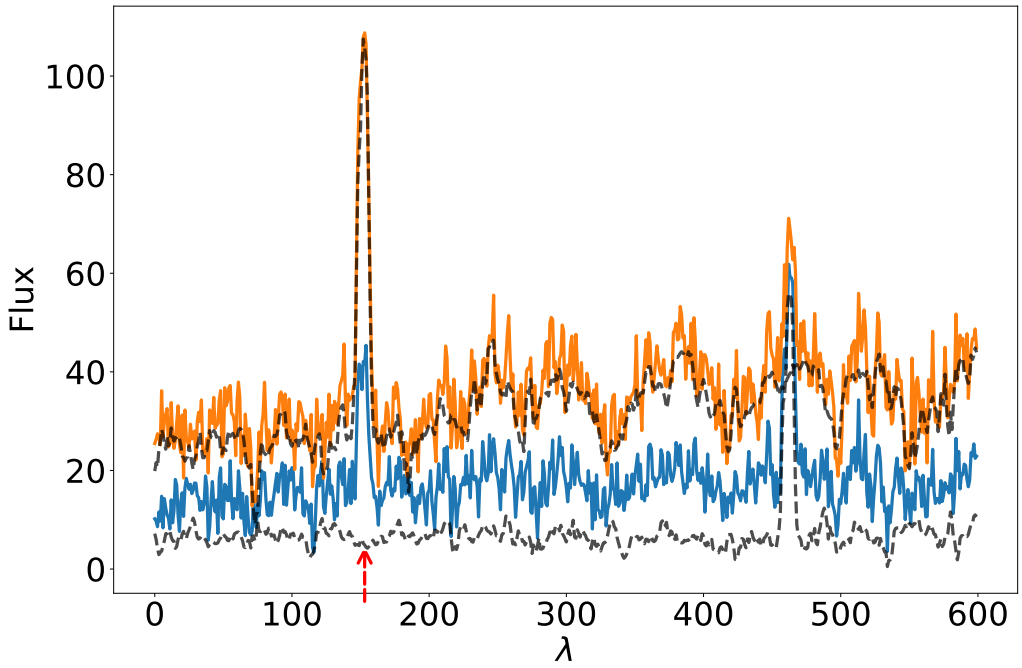
Les résultats présentés dans la figure 3.9 décrivent l'évolution des performances lorsque le conditionnement se dégrade. On peut voir que les performances se dégradent fortement lorsque le problème devient de plus en plus mal conditionné. En particulier, la variance des spectres estimés augmente et ces spectres deviennent fortement corrélés négativement. Cette anti-corrélation s'explique par la positivité de la matrice d'intensité $\widetilde{\boldsymbol{U}}$. En effet, si on se restreint à deux sources et qu'on note σ^2 la variance du bruit, alors la matrice de covariance de l'estimateur est $\sigma^2(\widetilde{\boldsymbol{U}}^T\widetilde{\boldsymbol{U}})^{-1}$, de dimension 2×2. Son terme hors-diagonal (terme de covariance) est alors toujours négatif [1]. Ce phénomène est illustré également par la figure 3.10 où sont montrés les spectres démélangés lorsque les sources sont très proches spatialement ($c \approx 4$). On peut voir sur cette figure que des artefacts positifs sur un spectre et négatifs sur l'autre apparaissent, se compensant et pouvant être de taille arbitrairement grande. Cela n'est absolument pas satisfaisant d'un point de vue applicatif car la variance des spectres estimés peut devenir comparable, voire supérieure, à l'énergie du signal d'intérêt. Les artefacts produits peuvent ainsi être interprétés à tort comme des raies d'émission ou d'absorption.

Nous allons donc désormais chercher à régulariser le problème afin de faire face à ces cas mal-conditionnés.

1. puisque $(\widetilde{\boldsymbol{U}}^T\widetilde{\boldsymbol{U}})^{-1} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}^{-1} = \frac{1}{\det(\widetilde{\boldsymbol{U}}^T\widetilde{\boldsymbol{U}})}\begin{bmatrix} c & -b \\ -b & a \end{bmatrix}$, avec a, b, c et $\det(\widetilde{\boldsymbol{U}}^T\widetilde{\boldsymbol{U}})$ supérieurs à zéro.

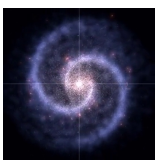


(a) Spectres de références vs spectres estimés par démélange.



(b) Spectres de références vs spectres estimés par simple somme sur une ouverture. La contamination d'une raie sur le second spectre est indiquée par une flèche rouge.

FIGURE 3.6 – Comparaisons entre les spectres de références et des spectres estimés par la méthode proposée et par une méthode directe de somme sur une ouverture spatiale. Les spectres de références sont en pointillés noirs. On peut voir que la méthode directe entraîne une contamination d'un premier spectre par le deuxième.



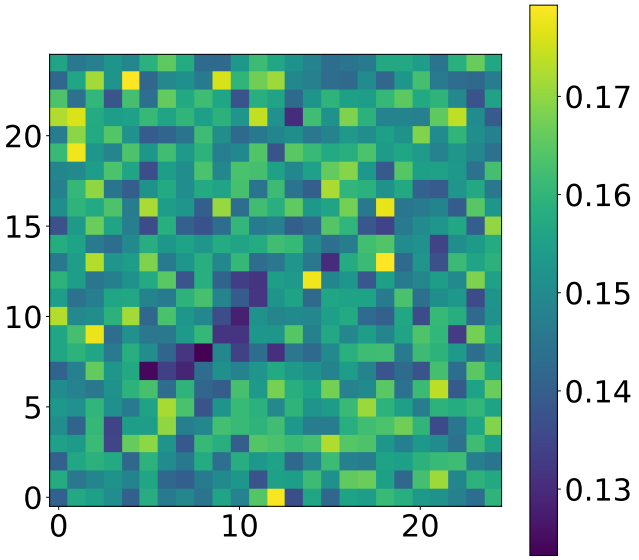


FIGURE 3.7 – Variance des résidus (le long de l'axe spectral), en unité arbitraire. La variance du bruit est 0.155.

3.3 Régularisation

Comme vu précédemment, lorsque la matrice d'intensité est mal conditionnée, c'est à dire lorsque des objets sont spatialement très proches, la résolution du système au sens des moindres carrés risque de sur-ajuster le bruit. Cela peut notamment se traduire par la création d'artefacts spectraux tels qu'illustrés en figure 3.10(b).

Afin de faire face à ce phénomène, une approche classique consiste à ajouter des contraintes afin de régulariser le problème. Il existe un grand nombre de régularisations possibles dans la littérature, parmi lesquelles on peut notamment citer les régularisations par pénalisation de type ridge (HOERL et KENNARD 1970) ou LASSO (TIBSHIRANI 1996) et les régularisations par critère informationnel comme l'*Akaike information criterion* (AIC) [AKAIKE 1974] ou le *Bayesian information criterion* (BIC) [SCHWARZ 1978]. Rappelons que l'ajout d'une contrainte de non-négativité des spectres, qui semble justifiable d'un point de vue physique (spectre mesurant un flux de photons), permettrait également de contribuer à régulariser le problème mais n'est pas satisfaisante ici notamment du fait des prétraitements de soustraction du fond de ciel.

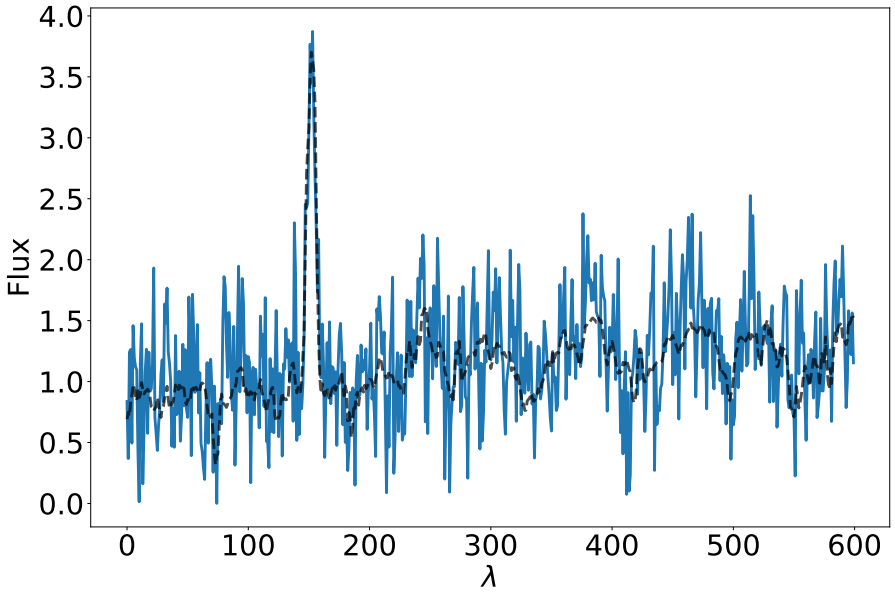
3.3.1 Régularisation par pénalisation

3.3.1.1 Régularisation ridge

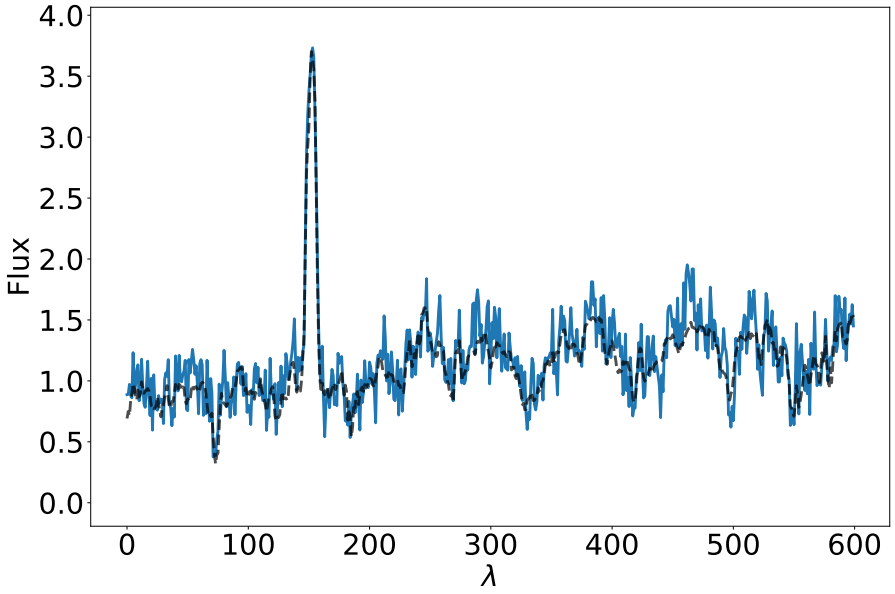
On s'intéresse ici à l'estimation de l'intensité spectrale des objets pour un bloc de feuillets donné l et une image HST i. Pour simplifier les notations, on note alors le vecteur colonne $\boldsymbol{y} = \underbrace{\boldsymbol{Y}_l}_{n\times 1}$, $\widetilde{\boldsymbol{U}} = \widetilde{\boldsymbol{U}}_{i,l}$ et $\boldsymbol{d} = \underbrace{\boldsymbol{D}_{i,l}}_{k\times 1}$. On a alors

$$\boldsymbol{y} = \widetilde{\boldsymbol{U}}\boldsymbol{d} + \boldsymbol{\epsilon}, \tag{3.13}$$

avec $\mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}$ et on peut supposer sans perte de généralité que $\mathbb{E}[\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}] = \sigma^2\boldsymbol{I}_n$. La régularisation ridge (HOERL et KENNARD 1970), qui est un cas particulier de la régularisation Tikhonov

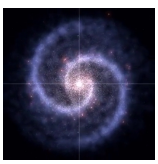


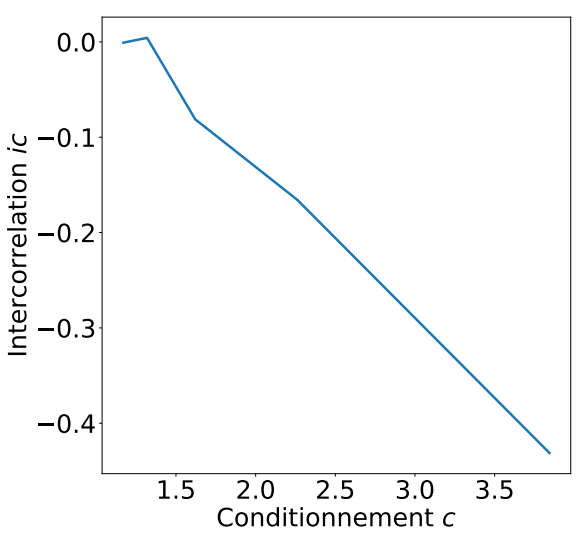
(a) Spectre de référence vs spectre bruité au centre de la source 2



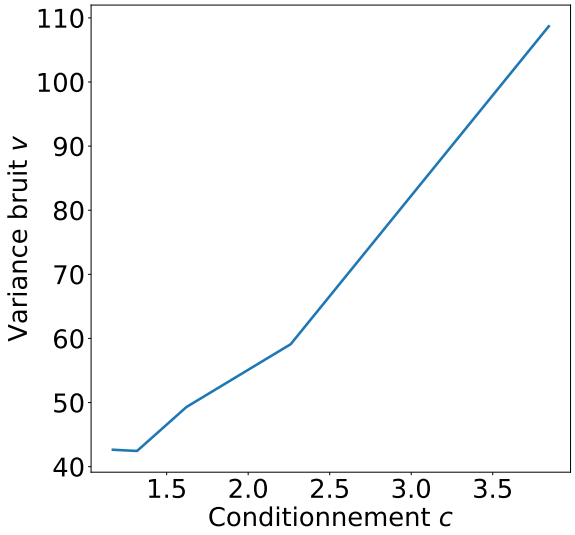
(b) Spectre de référence vs spectre estimé par démélange au centre de la source 2

FIGURE 3.8 – Comparaisons entre le spectre de référence, le spectre bruité et le spectre estimé par la méthode proposée, au centre de la source 2. Les spectres de références sont en pontillés noirs.





(a) Evolution de l'intercorrélation entre les 2 spectres estimés en fonction du conditionnement



(b) Evolution de la variance moyenne du bruit dans les 2 spectres estimés en fonction du conditionnement

FIGURE 3.9 – Evolution des performances en fonction du conditionnement de la matrice $\widetilde{\boldsymbol{U}}$.

[TIKHONOV et al. 1977], consiste à pénaliser le critère des moindres carrés ordinaires (MCO) par la norme ℓ_2 du vecteur $\boldsymbol{d}$ recherché. On cherche alors une estimation sous la forme

$$\widehat{\boldsymbol{d}} = \underset{\boldsymbol{d}}{\operatorname{argmin}} \, ||\boldsymbol{y} - \widetilde{\boldsymbol{U}}\boldsymbol{d}||_2^2 + \alpha||\boldsymbol{d}||_2^2, \tag{3.14}$$

où $\alpha \geq 0$ est le paramètre de régularisation à fixer. Ce problème admet une solution analytique de la forme

$$\widehat{\boldsymbol{d}} = \left(\widetilde{\boldsymbol{U}}^T\widetilde{\boldsymbol{U}} + \alpha\boldsymbol{I}_k\right)^{-1}\widetilde{\boldsymbol{U}}^T\boldsymbol{y}, \tag{3.15}$$

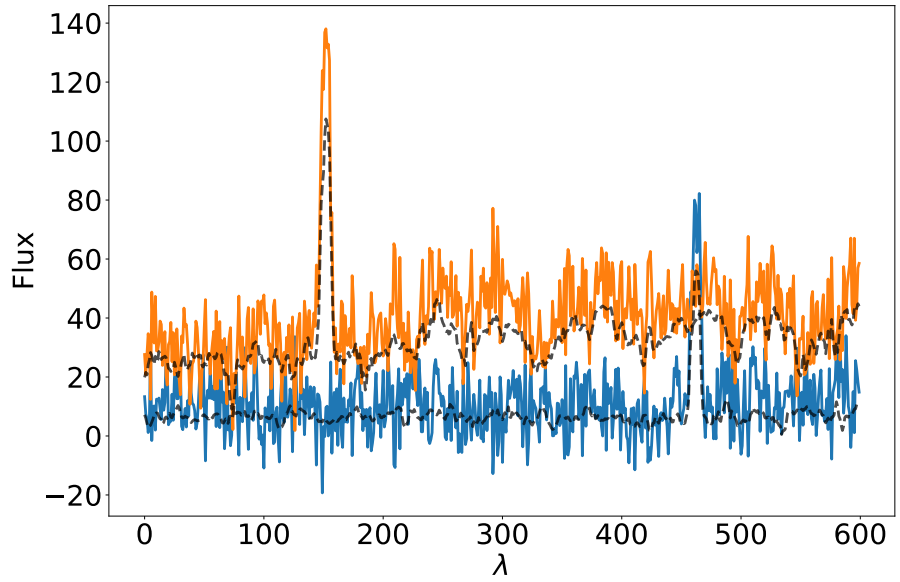
avec $\boldsymbol{I}_k$ la matrice identité de $\mathbb{R}^{k\times k}$ La formulation (3.14) peut être vue comme l'expression lagrangienne du problème d'optimisation sous contrainte (3.16).

$$\widehat{\boldsymbol{d}} = \underset{\boldsymbol{d}}{\operatorname{argmin}} \, ||\boldsymbol{y} - \widetilde{\boldsymbol{U}}\boldsymbol{d}||_2^2 \text{ avec } ||\boldsymbol{d}||_2 \leq t, \tag{3.16}$$

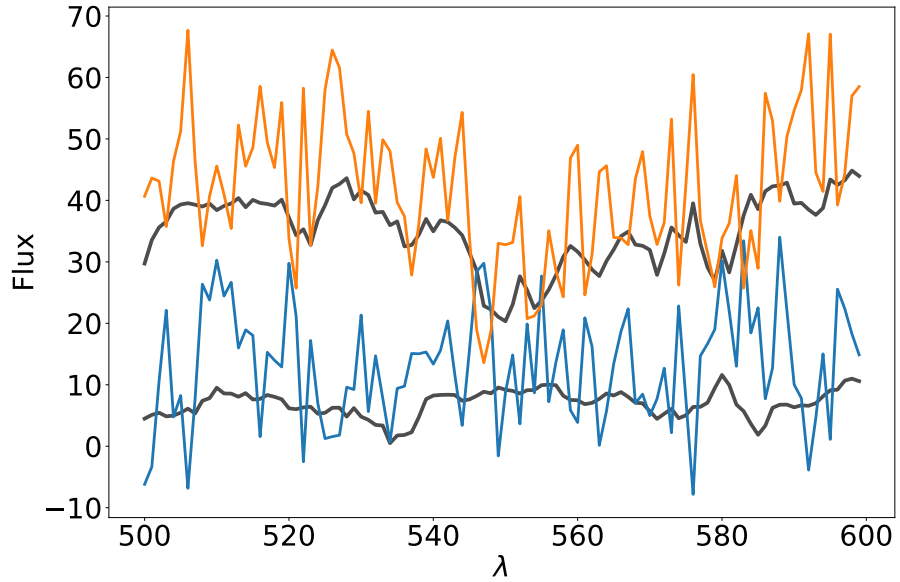
avec $t \geq 0$ fixé. En effet pour chaque valeur de $t \geq 0$, on peut trouver un $\alpha \geq 0$ tel que les solutions des problèmes (3.14) et (3.16) sont identiques. Cette dernière expression, appelée la forme contrainte du problème, montre qu'on cherche à limiter l'amplitude des composantes du vecteur $\boldsymbol{d}$ estimé.

L'ajout de ce terme de pénalité introduit un biais mais permet souvent de réduire fortement la variance de l'estimateur : rappelons en effet qu'il est toujours possible d'obtenir un terme multiplicateur α tel que l'erreur quadratique moyenne (EQM), qui correspond à la somme {variance+biais^2}, de l'estimateur régularisé soit inférieure à l'EQM de l'estimateur des moindres carrés [HOERL et KENNARD 1970 ; THEOBALD 1974]. L'EQM d'un estimateur $\widehat{\boldsymbol{d}}$ est en effet définie par

$$\begin{aligned} EQM(\widehat{\boldsymbol{d}}) &= \mathbb{E}[||\boldsymbol{d} - \widehat{\boldsymbol{d}}||_2^2] \\ &= \underbrace{\mathrm{Tr}(\mathrm{var}(\widehat{\boldsymbol{d}}))}_{variance} + \underbrace{||\boldsymbol{d} - \mathbb{E}[\widehat{\boldsymbol{d}}]||_2^2}_{biais^2} \end{aligned}$$



(a) Spectres estimés vs spectres de références



(b) Spectres estimés vs spectres de références (zoom sur une partie du spectre)

FIGURE 3.10 – Comparaisons entre les spectres de référence, et les spectres estimés au sens des moindres carrés. On peut voir l'apparition d'artefacts positifs/négatifs anticorrelés sur les spectres estimés.

Dans le cadre de l'estimateur des moindres carrés ordinaires, on a

$$EQM(\widehat{\boldsymbol{d}}_{LS}) = \sigma^2 \operatorname{Tr}\left[\left(\widetilde{\boldsymbol{U}}^T\widetilde{\boldsymbol{U}}\right)^{-1}\right]$$

(cf Hoerl et Kennard 1970), alors que l'EQM de l'estimateur α-régularisé vaut

$$EQM(\widehat{\boldsymbol{d}}_{\alpha}) = \underbrace{\sigma^2 \operatorname{Tr}\left(\boldsymbol{W}_{\alpha}\left(\widetilde{\boldsymbol{U}}^T\widetilde{\boldsymbol{U}}\right)^{-1}\boldsymbol{W}_{\alpha}^T\right)}_{variance} + \underbrace{\boldsymbol{d}^T\left(\boldsymbol{W}_{\alpha}-\boldsymbol{I}_k\right)^T\left(\boldsymbol{W}_{\alpha}-\boldsymbol{I}_k\right)\boldsymbol{d}}_{biais^2},$$

avec $\boldsymbol{W}_{\alpha} = \left[\boldsymbol{I}_k + \alpha(\widetilde{\boldsymbol{U}}^T\widetilde{\boldsymbol{U}})^{-1}\right]^{-1}$.

Le paramètre α permet de régler ce compromis biais-variance. Si $\alpha \to 0$, alors $\boldsymbol{W}_{\alpha} \to \boldsymbol{I}_k$, on retrouve l'estimateur au sens des moindres carrés avec un biais nul et la variance de l'estimateur des moindres carrés, potentiellement très grande si $\widetilde{\boldsymbol{U}}$ est mal conditionnée. Si $\alpha \to \infty$ on ne prend plus en compte le terme d'attache aux données et l'estimateur tend vers 0, ce qui correspond à une variance nulle et un biais égal à $\boldsymbol{d}^T\boldsymbol{d}$.

La figure 3.11 illustre le compromis biais-variance pour un exemple-jouet ($\widetilde{\boldsymbol{U}} \in \mathbb{R}^{20\times5}$ tiré selon une loi uniforme, $\boldsymbol{d} \in \mathbb{R}^{5\times1}$ tiré selon une loi normale, bruit gaussien). On peut voir que l'estimateur MCO ($\alpha = 0$) a bien un biais nul mais son EQM n'est pas l'EQM minimale, qui est obtenue pour $\alpha \approx 1$.

Notons également que la régularisation ridge possède un lien fort avec les approches de réduction de dimension par Analyse en Composantes Principales (ACP). En effet, si on note $\{s_j\}$ les valeurs singulières de $\boldsymbol{d}$ et qu'on se place, sans perte de généralité [2], dans le cas où $\boldsymbol{U}^T\boldsymbol{U} = \operatorname{diag}(\{s_j^2\})$ on a alors

$$\widehat{d_j}(\alpha) = \frac{s_j^2}{s_j^2+\alpha}\widehat{d}_j^{LS}$$

où $\widehat{\boldsymbol{d}}^{LS}$ est l'estimateur des moindres carrés et $\widehat{\boldsymbol{d}}(\alpha)$ est l'estimateur pour une régularisation ridge de paramètre α. On peut voir que les composantes associées aux grandes valeurs singulières sont peu affectées alors que les composantes associées aux faibles valeurs singulières sont fortement diminuées vers 0. On peut donc assimiler la régularisation ridge à une ACP assouplie.

3.3.1.2 Régularisation LASSO

Une autre approche de régularisation consiste à exploiter une hypothèse de parcimonie des paramètres recherchés. Afin d'éviter le sur-ajustement, on cherche à minimiser le nombre de prédicteurs (ici les spectres de sources) nécessaires pour expliquer les données. L'approche LASSO (*Least Absolute Shrinkage and Selection Operator*) [Tibshirani 1996] est une méthode de régularisation parcimonieuse et de sélection de variables, devenue très populaire depuis quelques années. Elle consiste à résoudre

$$\widehat{\boldsymbol{d}} = \underset{\boldsymbol{d}}{\operatorname{argmin}} \frac{1}{2} \times ||\boldsymbol{y} - \widetilde{\boldsymbol{U}}\boldsymbol{d}||_2^2 + \alpha \times ||\boldsymbol{d}||_1, \tag{3.17}$$

2. En effet on peut toujours réduire le problème de régression linéaire (3.13) à une forme canonique où la matrice $\boldsymbol{U}^T\boldsymbol{U}$ est diagonale, voir [Hoerl et Kennard 1970].

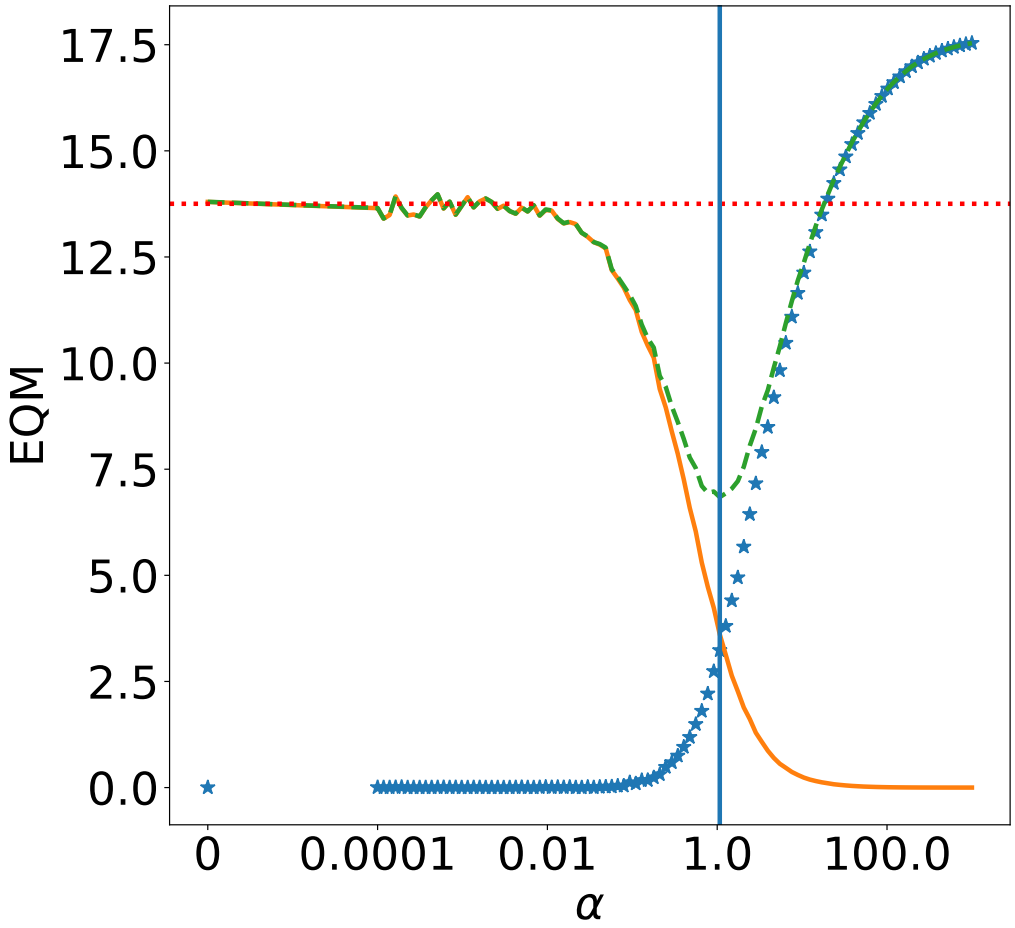
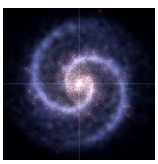


FIGURE 3.11 – Evolution de l'erreur quadratique moyenne (en pointillé vert) en fonction du paramètre de régularisation α pour un exemple-jouet ($\widetilde{\boldsymbol{U}} \in \mathbb{R}^{20\times 5}$ dont les entrées sont tirées selon une loi uniforme, $\boldsymbol{d} \in \mathbb{R}^{5\times 1}$, tiré selon une loi normale, bruit gaussien). En trait plein orange, le terme de variance et en $\star$ le terme de biais^2. L'EQM de l'estimateur MCO est indiqué en pointillé rouge et la valeur de régularisation α qui minimise l'EQM de l'estimateur régularisé est indiqué par une ligne verticale bleue. Les résultats sont obtenus après 5000 simulations Monte-Carlo.



ou de façon équivalente (cf ridge) sa formulation sous contrainte

$$\widehat{\boldsymbol{d}} = \underset{\boldsymbol{d}}{\operatorname{argmin}} ||\boldsymbol{y} - \widetilde{\boldsymbol{U}}\boldsymbol{d}||_2^2 \text{ avec } ||\boldsymbol{d}||_1 \leq t,$$

pour t fixé. Le choix de la norme ℓ_1 au détriment de la norme ℓ_2 de la régression ridge, permet d'améliorer l'interprétabilité des résultats en limitant le nombre de coefficients non-nuls : la norme ℓ_1 peut être vue comme une relaxation de la pseudo-norme ℓ_0 comptabilisant le nombre de coefficients non-nuls (Ramirez et al. 2013). La solution obtenue est donc en général parcimonieuse. Il n'existe toutefois pas de solution analytique à l'équation (3.17) mais des algorithmes performants comme le *least angle regression* (LARS) [Efron et al. 2004] , ou des des méthodes de descente de gradient par coordonnées (Friedman et al. 2010), permettent de résoudre numériquement de façon efficace ce problème.

Le group-LASSO (Yuan et Lin 2006) permet d'améliorer la robustesse des solutions estimées en sélectionnant les mêmes variables (sources) pour un ensemble de tâches (longueurs d'onde). Au lieu de travailler sur un seul feuillet spectral, on travaille sur un bloc spectral l de taille λ_l sur lequel $\widetilde{\boldsymbol{U}}$ est supposée constante : on a $\boldsymbol{Y}_l \in \mathbb{R}^{n\times\lambda_l}$, et on cherche donc à estimer les spectres $\boldsymbol{D}_l \in \mathbb{R}^{k\times\lambda_l}$ des sources sur ce domaine spectral. La fonction de coût devient

$$\frac{1}{2} \times ||\boldsymbol{Y} - \widetilde{\boldsymbol{U}}\boldsymbol{D}||_2^2 + \alpha||\boldsymbol{D}||_{21}, \tag{3.18}$$

où $||\boldsymbol{D}||_{21} = \sum_i \sqrt{\sum_j d_{ij}^2}$. Cette norme mixte permet d'imposer une parcimonie "par groupe de feuillets", autrement dit sur les lignes de la matrice $\boldsymbol{D}$.

3.3.2 Régularisation par critère informationnel

La prise en compte d'une contrainte de parcimonie peut également se faire par une approche de sélection de modèles à l'aide d'un critère de décision comme le Bayesian Information Criterion (BIC) ou l'Akaike Information Criterion (AIC).

3.3.2.1 Bayesian Information Criterion (BIC)

Les approches de sélection de modèles peuvent se faire notamment à l'aide du Bayesian Information Criterion (BIC) [Schwarz 1978]. Le critère BIC s'écrit $\text{BIC} = K\log(n) - 2\log(\hat{L})$ où $\hat{L}$ est le maximum de vraisemblance du modèle, K est le nombre de paramètres libres à estimer, n est le nombre d'échantillons (ici le nombre de pixels dans la zone étudiée). Dans notre cas, on fait l'approximation que le bruit est gaussien (cf chapitre 1). Si on travaille sur un feuillet spectral, pour un modèle $\mathcal{M}$ avec k objets à spectre non-nul, on a alors :

$$\text{BIC}(\mathcal{M}) = K\log(n) + \log(\widehat{\sigma_{\mathcal{M}}}^2) \tag{3.19}$$

où $\widehat{\sigma_{\mathcal{M}}}^2$ est la variance empirique des résidus obtenus en utilisant le modèle $\mathcal{M}$ choisi. On a un paramètre libre par spectre non nul plus un pour la variance du bruit d'où $K = k+1$. Pour minimiser le critère BIC on voit qu'il y a un compromis à trouver entre un modèle expliquant au mieux les données ($\widehat{\sigma}^2$ minimal) et un modèle le plus simple possible (k minimal).

La sélection du meilleur modèle se ramène alors à un problème de combinatoire où on teste toutes les combinaisons possibles de spectres non-nuls et de conserver celle qui minimise le critère BIC.

3.3.2.2 Akaike Information Criterion (AIC)

Le critère AIC (Akaike 1973) est en apparence très semblable au critère BIC puisqu'il s'écrit $\text{AIC} = 2K - 2\log(\hat{L})$. La seule différence se situe dans le terme évaluant la complexité du modèle, qui ne dépend plus du nombre d'échantillons comme le BIC mais uniquement du nombre de degrés de liberté K. Conceptuellement le critère AIC est toutefois très différent du BIC puisqu'il cherche avant tout à minimiser l'erreur de prédiction et à être asymptotiquement efficace (Aho et al. 2014).

3.3.2.3 Comparaison des critères

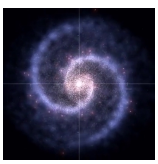
Comme cela apparaît clairement dans l'expression des deux critères, le BIC sera plus agressif que l'AIC dès lors que $\log(n) \geq 2$, et ce d'autant plus que le nombre d'échantillons n augmente. Les différences conceptuelles entre les deux critères mènent à des propriétés asymptotiques différentes. Lorsque le nombre d'échantillons grandit, la sélection selon AIC devient équivalente à une sélection basée sur le maximum de vraisemblance, puisque le terme prenant en compte les degrés de liberté devient négligeable. Au contraire le BIC ne devient pas équivalent au maximum de vraisemblance du fait de la présence de n dans le terme pénalisant la complexité du modèle. Le BIC est donc consistent pour une sélection d'un modèle au sein d'une famille de modèle (contenant le vrai modèle) [Poskitt 1987 ; Haughton 1988], alors que l'AIC favorise des modèles de plus en plus complexes lorsque le nombre de données augmente. On peut à nouveau voir ce choix comme un compromis biais-variance.

3.3.3 Stratégie choisie

Afin de répondre aux spécificités des données MUSE, nous avons choisi d'exploiter de façon complémentaire les régularisations parcimonieuses et ridge. En effet les spectres recherchés sont composés d'un continuum spectral (ou ligne de base) à variation lente et d'un ensemble de raies d'émission et d'absorption. Ces raies sont a priori réparties de façon parcimonieuse parmi les objets/galaxies, c'est à dire qu'à une longueur d'onde donnée, seul un très petit nombre de sources possèdent une raie. Par ailleurs, la plupart des galaxies possèdent également un continuum spectral présent sur toutes les longueurs d'onde. Du fait des dynamiques d'intensité très différentes entre le continuum et les raies, appliquer globalement une régularisation (par exemple ridge) sur les spectres à estimer ne s'avère pas suffisamment robuste pour être appliquer de manière non-supervisée (i.e. sans réglage manuel des paramètres de régularisation) sur les données réelles.

Nous proposons donc de traiter séparément les raies et le continuum. Pour cela la démarche est la suivante :

1. Obtention du cube des raies $\boldsymbol{Y}^r$ par soustraction du cube de continuum estimé par un filtrage robuste (voir annexe B).
2. Reconstruction des raies $\widehat{\boldsymbol{D}}^r$ associées aux objets à partir du cube des raies obtenu précédemment : nous verrons qu'une approche de régularisation parcimonieuse avec sélection de modèles a été choisie.
3. Soustraction de la contribution estimée des raies pour obtenir les données restantes $\boldsymbol{Y}^c = \boldsymbol{Y} - \widetilde{\boldsymbol{U}}\widehat{\boldsymbol{D}}^r$



4. Estimation des continuums spectraux $\widehat{\boldsymbol{D}}^c$ à l'aide d'une régularisation ridge sur les données $\boldsymbol{Y}^c$
5. Combinaison des deux estimations $\widehat{\boldsymbol{D}} = \widehat{\boldsymbol{D}}^r + \widehat{\boldsymbol{D}}^c$

3.3.4 Reconstruction des raies

L'intérêt d'une régularisation parcimonieuse pour les raies permet non seulement d'éviter les phénomènes de sur-ajustement du bruit, mais également d'encourager l'attribution d'une raie à un nombre minimal de spectres, ce qui est physiquement le plus probable.

On sait qu'on recherche des raies qui s'étendent sur plusieurs feuillets spectraux. On peut donc exploiter cet a priori pour assurer le plus possible une bonne reconstruction des raies et éviter des discontinuités spectrales (où une partie de la raie serait associé à un objet et une autre partie à un autre objet). Pour cela la sélection de modèles peut se faire conjointement sur tout le support spectral de la raie. Il est donc nécessaire d'estimer ce support spectral.

La démarche sera donc la suivante :

- Détection rapide de tous les supports potentiels de raies spectrales.
- Sur chacun des supports, sélection des objets à spectre non-nul, puis estimation des spectres en ne prenant en compte que les objets sélectionnés.

3.3.4.1 Détection des supports de raie

La détection des supports de raie se fait en s'appuyant sur les a priori physiques en notre possession. Afin de limiter le nombre de données à explorer on travaille sur un ensemble de spectres estimés $\boldsymbol{D}_0$: on somme pour cela les données pondérées par la carte d'intensité de chaque objet pour obtenir un spectre par objet) :

$$\boldsymbol{D}_0 = \widetilde{\boldsymbol{U}}^T \boldsymbol{Y} \tag{3.20}$$

Sur chacun des spectres $\boldsymbol{d}$ (lignes de $\boldsymbol{D}_0$), on cherche des pics ayant une taille minimale de quelques feuillets et une forme gaussienne en première approximation. Notons $\boldsymbol{g}_m$ de taille L un vecteur représentant un noyau gaussien tronqué, centré en $L/2$ et de variance $\frac{w}{2}^2$. La largeur caractéristique m de ce noyau est définie ici comme correspondant à quatre fois l'écart-type de la densité gaussienne). Un tel noyau est présenté en figure 3.12. On note respectivement w et w_{max} la largeur minimale et la valeur maximale acceptées.

On commence par effectuer un filtrage adapté avec le noyau gaussien $\boldsymbol{g}_w$ correspondant à la largeur minimale fixée w.

$$\widetilde{\boldsymbol{d}} = \boldsymbol{d} * \boldsymbol{g}_w \tag{3.21}$$

On détecte ensuite l'ensemble $\mathcal{S}_0$ des extrema locaux espacés de cette largeur minimale w :

$$\mathcal{S}_0 = \{T \text{ tel que } \widetilde{d}_T = \max_{T-w \leq t \leq T+w} \widetilde{d}_t\} \tag{3.22}$$

Sous l'hypothèse de bruit gaussien, on peut contrôler le taux de fausses alarmes grâce à une estimation simple de la variance du bruit par le *Median Absolute Deviation* (MAD)[Rousseeuw et Croux 1993]. En effet, sous l'hypothèse de bruit gaussien, un estimateur robuste de l'écart-type σ du vecteur $\widetilde{\boldsymbol{d}}$ est :

$$\widehat{\sigma}_{\text{MAD}} = \frac{1}{\Phi^{-1}(3/4)} \text{MAD}(\widetilde{\boldsymbol{d}}), \tag{3.23}$$

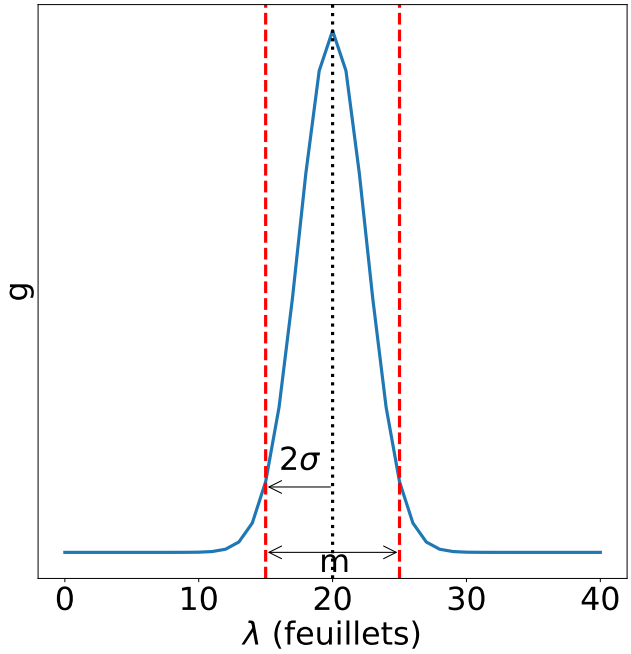


FIGURE 3.12 – Exemple de vecteur de noyau gaussien de taille $m = 10$ (la densité gaussienne a donc pour écart-type $\sigma = 2.5$)

où $\text{MAD}(\widetilde{\boldsymbol{d}}) = \text{mediane}(|\widetilde{\boldsymbol{d}} - \text{mediane}(\widetilde{\boldsymbol{d}})|)$ et Φ^{-1} est la fonction quantile de la distribution normale. On conserve alors uniquement l'ensemble $\mathcal{S}_1$ des extrema supérieurs à $a \times \widehat{\sigma}_{\text{MAD}}$ où a est fixé en fonction du nombre moyen de fausses alarmes autorisé :

$$\mathcal{S}_1 = \{T \in \mathcal{S}_0 \text{ tel que } \widetilde{d}_T =\leq a\widehat{\sigma}_{\text{MAD}}\} \tag{3.24}$$

On rejette ensuite également tous les extrema qui ne possèdent aucune continuité spectrale (sur les spectres non filtrés) par une règle simple : les voisins spectraux doivent être du même signe et d'amplitude supérieure à un écart-type. On aboutit alors à l'ensemble $\mathcal{S}_2$:

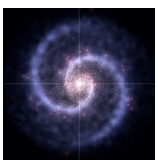
$$\mathcal{S}_2 = \{T \in \mathcal{S}_1 \text{ tel que } \forall -1 \leq i \leq 1 \quad d_{T+i} \geq \text{sgn}(d_T)\widehat{\sigma}_{\text{MAD}}\} \tag{3.25}$$

Enfin on détermine la taille de chaque support en cherchant le maximum de corrélation avec une famille de noyaux gaussiens $\mathcal{G}$ de largeur variable : $\mathcal{G} = \{\boldsymbol{g}_j\}$ où $\boldsymbol{g}_j \in \mathbb{R}^{w_{max}}$ est un noyau gaussien de largeur caractéristique m_j, avec $w \leq m_j \leq w_{max}$ (les vecteurs sont tous de taille w_{max} pour permettre une comparaison pertinente). Pour chaque extremum l conservé on tronque le spectre $\boldsymbol{d}$ en un vecteur $\boldsymbol{d}_l \in \mathbb{R}^{w_{max}}$ centré sur l'extremum. On obtient alors la largeur caractéristique w_l de la raie

$$w_l = \underset{j}{\text{argmax}} \frac{\boldsymbol{d}_l^T \boldsymbol{g}_j}{||\boldsymbol{d}_l||.||\boldsymbol{g}_j||} \tag{3.26}$$

On peut alors construire l'ensemble des largeurs des raies $\mathcal{W}$.

Cette routine est résumée dans l'algorithme 2. Elle permet d'obtenir rapidement l'ensemble des supports spectraux des raies potentielles dans la zone étudiée. On peut alors faire une estimation des spectres des objets avec régularisation par sélection de modèles sur chacun de ces supports.



Algorithme 2 Procédure de détection des supports de raies

Entrées : matrice de raies $\boldsymbol{Y}_l$, largeur minimale w, largeur maximale w_{max}, coeff a

$\mathcal{S} = \emptyset$, $\mathcal{W} = \emptyset$ ▷ Initialisation
$\boldsymbol{D}_0 = \boldsymbol{U}^T \boldsymbol{Y}_l$ ▷ Construction des spectres à explorer
for spectre $\boldsymbol{d}_i$ **do**
    Calcul de $\widetilde{\boldsymbol{d}}$ avec l'eq. (3.21) ▷ Filtrage adapté par un noyau gaussien
    Calcul de $\mathcal{S}_0$ par l'eq. (3.22) ▷ Détection des extrema locaux
    Calcul de $\mathcal{S}_1$ par l'eq. (3.24) ▷ Rejet des extrema locaux trop faibles
    Calcul de $\mathcal{S}_2$ par l'eq. (3.25) ▷ Rejet des extrema locaux sans continuité spectrale
    $\mathcal{S} = \mathcal{S} \cup \mathcal{S}_2$
    for extremum l **do**
        Calcul de w_l avec l'eq. (3.26) ▷ Estimation de la largeur de la raie
        $\mathcal{W} = \mathcal{W} \cup \{w_l\}$

Sorties : $\mathcal{S}$ et $\mathcal{W}$ ▷ Liste des raies et leurs largeurs

Les différents paramètres ont été calibrés manuellement sur données simulées avant d'être validés sur quelques exemples issus des données réelles. Il a notamment été choisi de fixer $w = 3$, $w_{max} = 30$ et le paramètre $a = 2.5$ ce qui correspond à un nombre moyen de fausses alarmes proche de un pour les 3600 bandes spectrales d'un spectre. L'objectif est de limiter le nombre de faux positifs pour ne pas devoir tester un nombre trop important de raies. Il est à noter que ce processus de détection ne cherche pas nécessairement à être exhaustif à tout prix mais a seulement pour but d'optimiser la prise en compte de l'information physique de parcimonie des raies. Si certaines raies ne sont pas détectées, le signal correspondant sera alors simplement exploité lors de l'estimation des continuums spectraux, sans a priori de parcimonie.

On peut voir sur la figure 3.13 un exemple simulé de détection de raies à partir des spectres estimés $\boldsymbol{D}_0$ par l'équation (3.20).

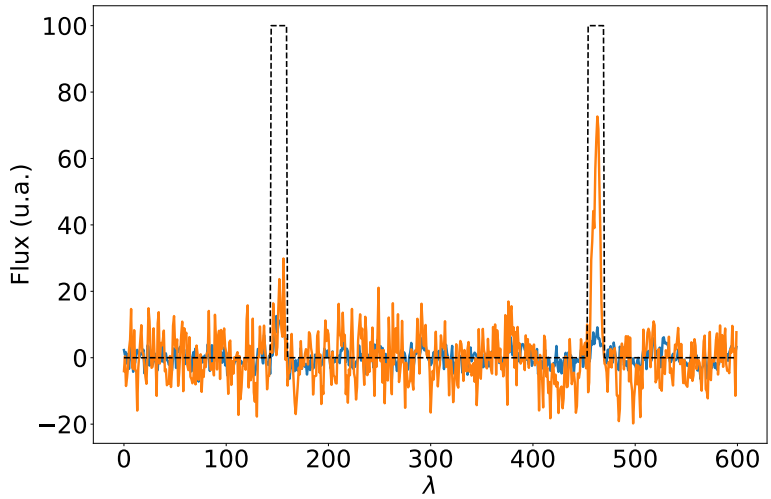


FIGURE 3.13 – Détection des raies sur des estimations simples des spectres des objets. En pointillé noir les supports des raies détectées.

3.3.4.2 Sélection parcimonieuse et estimation des raies

Comme vu dans les paragraphes précédents, cette régularisation parcimonieuse peut être mise en place à l'aide d'un terme de pénalité ajouté au critère des moindres carrés (3.3) via l'approche group-LASSO, ou à l'aide d'un critère informationnel comme le BIC et l'AIC.

L'approche group-LASSO nécessite le choix du paramètre de régularisation α. On peut pour cela utiliser une approche de validation croisée afin d'estimer le paramètre α minimisant l'erreur quadratique moyenne. Par ailleurs, ce type de régularisation pénalisant l'intensité des coefficients recherchés entraîne par définition une perte importante du flux dans les spectres reconstruits. Pour limiter ce problème, on peut effectuer la régularisation en deux temps : on commence par estimer quels sont les coefficients non-nuls obtenus par le group-LASSO, puis on effectue une nouvelle résolution du problème au sens des moindres carrés (3.3) en ne prenant en compte que les objets sélectionnés par le group-LASSO. Pour que cela s'avère efficace, il est nécessaire que le group-LASSO soit suffisamment agressif en amenant le plus grand nombre de coefficients non pertinents à zéro. Pour cela, on peut utiliser la règle heuristique d'une erreur standard ("1SE") [Breiman et al. 1984] (détaillée dans le paragraphe 3.3.5.1).

L'autre possibilité pour imposer une régularisation parcimonieuse consiste donc à utiliser un critère informationnel comme le BIC et l'AIC. Dans le cadre de notre application, on souhaite surtout favoriser la parcimonie et associer une raie à un nombre minimal de sources. On s'attend ici à ce que le "vrai modèle" (la configuration de sources possédant la raie) soit bien présent dans les modèles testés. Rappelons également qu'on peut se permettre de choisir un critère de sélection agressif sans trop de risque car le signal qui ne sera pas expliqué par cette étape, sera pris en compte lors du démélange des continuums. Le critère BIC semble donc plus adapté à notre problème que le critère AIC.

Nous avons donc comparé l'approche group-LASSO, en utilisant une validation croisée et l'heuristique "1SE", et la sélection par critère BIC sur des données simulant notre application de démélange. Les deux approches donnent des résultats comparables, avec un plus faible coût calculatoire et une plus grande simplicité d'implémentation pour la sélection par critère informationnel qui ne nécessite pas d'estimation d'un paramètre de régularisation.

Il est à noter que ce type d'approche où l'estimation se fait après la sélection du modèle permet de fortement limiter le biais par rapport à un group-LASSO simple (c'est à dire non suivi d'une estimation MCO sur les spectres sélectionnés) [Belloni et Chernozhukov 2009]. Ce biais reste néanmoins non nul du fait de la sélection parcimonieuse.

Remarque 4

La sélection de modèles par critère BIC (ou AIC) nécessite de tester successivement tous les modèles possibles, c'est à dire toutes les combinaisons de spectres non-nuls. En pratique, lorsqu'on travaille sur tout un cube MUSE ou même sur des zones de mélanges dans les données réelles (où une vingtaine d'objets sont concentrés sur une zone de 30 par 30 pixels), le test exhaustif de toutes les combinaisons devient bien sûr impossible en un temps de calcul raisonnable. On procède alors à une variante gloutonne, où on ajoute itérativement au modèle, par ordre décroissant, l'objet dont la carte d'intensité est la plus corrélée aux résidus issus de la régression par le modèle précédent.

Remarque 5

Il est à noter également que l'approche BIC n'est plus consistante si on augmente fortement

le nombre m de feuillets considérés. En effet, on quitte son domaine de validité asymptotique (le nombre de paramètres devient trop important). L'expression du BIC pour un nombre m de feuillets devient en effet $\mathrm{BIC} = (k \times m + 1)\log(n \times m) + m\log(\widehat{\sigma}^2)$ *On voit alors qu'une solution avec un plus grand nombre de paramètres donnera toujours un critère BIC plus grand (car le terme dominant est en* $\log(n \times m) \times (k \times m + 1)$ *) et sera donc systématiquement rejetée. Afin d'éviter ce problème, la sélection de modèle se fait sur les données moyennées sur le support spectral. La régression pour estimer la forme des raies se fait sur les données originelles (non moyennées) avec le modèle sélectionné.*

Pour chaque support de raie l, on sélectionne donc le modèle $\mathcal{M}_l$ sur le feuillet moyen du bloc l à l'aide de la formule

$$\mathcal{M}_l = \underset{\mathcal{M}}{\operatorname{argmin}}\, \mathrm{BIC}(\mathcal{M}), \tag{3.27}$$

où $\mathrm{BIC}(\mathcal{M})$ est la valeur BIC du modèle $\mathcal{M}$ calculée par l'équation (3.19).

Si on note alors $\widetilde{\boldsymbol{U}}_{\mathcal{M}}$ la matrice des intensités contenant uniquement les colonnes des sources sélectionnées dans le modèle $\mathcal{M}$, on calcule finalement les spectres de raies associés $\widehat{\boldsymbol{D}}_l^r$ à la raie l par

$$\widehat{\boldsymbol{D}}_l^r = \underset{\boldsymbol{D}}{\operatorname{argmin}}\, ||\boldsymbol{Y}_l^r - \widetilde{\boldsymbol{U}}_{\mathcal{M}_l}\boldsymbol{D}||_2^2 \tag{3.28}$$

3.3.5 Régularisation du continuum

Après soustraction de la contribution estimée des raies, on cherche désormais à estimer les continuums spectraux des objets. Comme vu précédemment on ne peut plus ici exploiter d'hypothèse de parcimonie. La régularisation ridge permet alors de limiter le sur-ajustement du bruit pour un coût calculatoire équivalent à celui de l'estimation des moindres carrés.

3.3.5.1 Estimation du paramètre de régularisation

L'utilisation d'une régularisation ridge nécessite le choix crucial du paramètre de régularisation α. Un certain nombre de méthodes ont développées dans la littérature pour estimer la valeur la plus pertinente (Galatsanos et Katsaggelos 1992 ; Arlot et Celisse 2010). Une approche classique est d'estimer la valeur optimale (au seins de l'EQM) par validation croisée (CV) [Kohavi 1995]. Cela consiste à partitionner les données en plusieurs parties et, pour chaque paramètre de régularisation α, à estimer les paramètres sur une partie des données puis calculer l'erreur de prédiction sur les données restantes et de répéter ce processus. On choisit alors le paramètre α qui minimise l'erreur de prédiction moyennée sur tous les partitionnements de données. Un partitionnement possible est l'approche K-fold (Geisser 1975) qui consiste à séparer les données en K sous-ensembles de taille semblable. Chaque ensemble sert alors une fois d'ensemble de validation, tous les autres servant alors de données d'apprentissage. On effectue donc K estimations différentes. Le choix du nombre de sous-ensembles K a fait l'objet de nombreuses études empiriques, les valeurs classiquement retenues se situant autour de la dizaine (Kohavi 1995). Le cas extrême consiste à prendre K égal à n le nombre d'échantillons. On appelle cette stratégie, qui consiste à isoler successivement tous les échantillons un par un, le *Leave-One-Out* (LOO). L'apprentissage se fait sur tous les échantillons sauf un et l'erreur de prédiction est calculée sur l'échantillon restant.

Cette méthode possède un fort attrait dans le cadre d'une régularisation ridge car on peut exprimer très facilement les erreurs de prédiction en fonction des paramètres estimés une seule

fois sur l'ensemble des données. Notons tout d'abord qu'on peut obtenir à faible coût $\widehat{\boldsymbol{d}}$ pour plusieurs valeurs de α. En effet soit $\widehat{\boldsymbol{d}}_\alpha = \underset{\boldsymbol{d}}{\operatorname{argmin}} ||\boldsymbol{y} - \widetilde{\boldsymbol{U}}\boldsymbol{d}||_2^2 + \alpha||\boldsymbol{d}||_2^2$ et $\widetilde{\boldsymbol{U}} = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{V}^T$ la décomposition en valeurs singulières de $\widetilde{\boldsymbol{U}}$, avec $\boldsymbol{\Sigma} = \operatorname{diag}(s_j)$.

On a alors $\widehat{\boldsymbol{d}}(\alpha) = (\widetilde{\boldsymbol{U}}^T\widetilde{\boldsymbol{U}} + \alpha\boldsymbol{I})^{-1}\widetilde{\boldsymbol{U}}^T\boldsymbol{y}$ soit en utilisant la décomposition en valeurs singulières :

$$\widehat{\boldsymbol{d}}(\alpha) = \boldsymbol{V}\operatorname{diag}\left(\frac{s_j}{s_j^2 + \alpha}\right)\boldsymbol{A}^T\boldsymbol{y} \tag{3.29}$$

Obtenir $\widehat{\boldsymbol{d}}$ pour plusieurs valeurs de α revient alors simplement faire varier la matrice diagonale dans le produit matriciel (3.29). Pour un α donné, la validation croisée LOO consiste désormais à estimer $\boldsymbol{d}$ sur l'ensemble des échantillons/pixels sauf un pixel i. Notons $\widehat{\boldsymbol{d}}_{-i} = (\widetilde{\boldsymbol{U}}^T\widetilde{\boldsymbol{U}} + \alpha\boldsymbol{I})^{-1}_{-i}\widetilde{\boldsymbol{U}}^T_{-i}\boldsymbol{y}_{-i}$ cette estimation (par simplification on omet la dépendance en α). On calcule ensuite l'erreur de prédiction sur le pixel i, soit

$$cv_i = [\boldsymbol{y}_i - (\widetilde{\boldsymbol{U}}\widehat{\boldsymbol{d}}_{-i})_i]^2 \tag{3.30}$$

On peut obtenir facilement $\widehat{\boldsymbol{d}}_{-i}$ à partir de $\widehat{\boldsymbol{d}}$ à l'aide de la formule de mise à jour de Sherman-Morrison (Sherman et Morrison 1949) qui permet d'inverser facilement une matrice privée d'une ligne/colonne à partir de l'inverse de la matrice d'origine. L'erreur de prédiction moyenne s'écrit alors finalement

$$CV = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - (\widetilde{\boldsymbol{U}}\widehat{\boldsymbol{d}})_i}{1 - \boldsymbol{S}_{ii}}\right)^2$$

où $\boldsymbol{S} = \widetilde{\boldsymbol{U}}(\widetilde{\boldsymbol{U}}^T\widetilde{\boldsymbol{U}} + \alpha\boldsymbol{I})^{-1}\widetilde{\boldsymbol{U}}^T$. Notons une fois encore que $\boldsymbol{S}$ (et a fortiori ses seuls éléments diagonaux), peuvent être calculés sans inversion matricielle à l'aide de la décomposition en valeurs singulières de $\widetilde{\boldsymbol{U}}$.

Une alternative à la validation croisée LOO est la validation croisée généralisée (generalized cross validation ou GCV) [Golub et al. 1979], qui revient à approximer les éléments diagonaux de $\boldsymbol{S}$ à la valeur moyenne de la trace de $\boldsymbol{S}$. La GCV offre des propriétés d'invariance par rotation et permet dans notre cadre de diminuer le coût calculatoire. On obtient alors

$$GCV = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - (\widetilde{\boldsymbol{U}}\widehat{\boldsymbol{d}})_i}{1 - Tr(\boldsymbol{S})/n}\right)^2. \tag{3.31}$$

On a donc alors une erreur de prédiction en fonction du paramètre de régularisation α, comme illustré dans la figure 3.14. Plutôt que de prendre la valeur du minimum, on peut appliquer la règle heuristique de "$1SE$" (Breiman et al. 1984). En effet on peut voir que la courbe présente un plateau le long duquel l'erreur de prédiction varie très peu. Notons CV_m l'erreur de prédiction moyenne (sur l'ensemble des pixels) minimale (par rapport à α) obtenue pour $\alpha = \alpha_m$ et σ_{CV} l'écart-type des erreurs de prédictions :

$$\sigma_{CV}^2 = \frac{1}{n\sqrt{n}}\sum_{1\leq i\leq n} cv_i^2, \tag{3.32}$$

où cv_i est ici l'erreur de prédiction sur le pixel i pour α_m (le facteur $\sqrt{n}$ dans cette formule provient du fait qu'on cherche bien l'écart-type de la valeur moyennée CV_m). La règle "$1SE$"

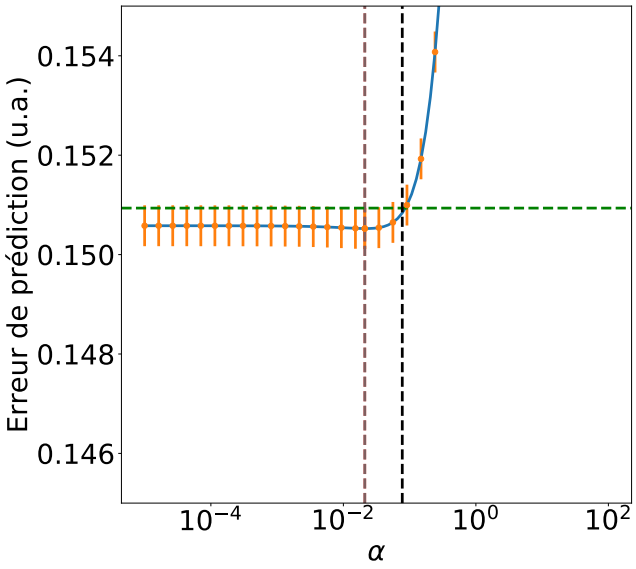


FIGURE 3.14 – Erreur de prédiction (en unité arbitraire) de la validation croisée en fonction du paramètre de régularisation α. Au lieu de sélectionner le α minimisant l'erreur (indiqué par les pointillés violet), on peut appliquer la règle "$1SE$" pour accentuer la régularisation : on cherche le α maximal (indiqué en pointillé noir) tel que l'erreur reste à $1SE$ (pointillés verts) de l'erreur minimale. Les écarts-type de l'erreur de prédiction moyenne sont indiqués en orange.

consiste alors à prendre le paramètre de régularisation α le plus régularisant possible tout en gardant une erreur de prédiction $CV \leq CV_m + \sigma_{CV}$:

$$\alpha_{m+1s} = \max_{\alpha}\{CV(\alpha) \leq CV_m + \sigma_{CV}\} \tag{3.33}$$

Le processus d'estimation du paramètre de régularisation α par GCV est décrit dans l'algorithme 3.

Algorithme 3 Procédure d'estimation du paramètre de régularisation α

1: *Entrée :* matrice de données $\boldsymbol{Y}$, matrice d'intensité $\widetilde{\boldsymbol{U}}$, liste de paramètres $\{\alpha\}$

2: **for** paramètre de régularisation α **do**
3: Calcul de $GCV(\alpha)$ à l'aide de (3.31) ▷ Calcul des erreurs de prédiction moyennes
4: Calcul de α_m ▷ Obtention du α minimisant l'erreur
5: Calcul des cv_i pour α_m avec l'éq. (3.30) ▷ Calcul des erreurs de prédiction pixeliques
6: Calcul de σ_{CV} avec l'equation 3.32 ▷ Calcul de la variance
7: Estimation de α_{m+1s} par (3.33) ▷ Utilisation par l'heuristique "$1SE$"
8: *Sortie :* α_{m+1s} ▷ Paramètre de régularisation choisi

Remarque 6

Afin d'accroître la robustesse de la validation croisée, un seul paramètre de régularisation est recherché par bloc de quelques dizaines de feuillets spectraux. Cela revient à considérer que le rapport signal-à-bruit des données (privées des raies) évolue faiblement en fonction de la longueur d'onde (rappelons que la matrice $\widetilde{\boldsymbol{U}}$ est elle aussi supposée constante sur le bloc spectral

considéré). Cette hypothèse est valable si on considère que les raies du ciel ont été suffisamment bien nettoyées.

3.3.5.2 Correction du flux

Un inconvénient naturel de la régularisation par ajout d'un terme de pénalité type ridge (ou LASSO) est que la solution estimée est biaisée vers zéro. Ce biais se traduit dans notre cas par une perte du flux sur les spectres estimés, ce qui n'est pas souhaité pour l'analyse astrophysique des spectres. Pour pallier à cet inconvénient, nous pouvons ici profiter de la redondance des estimateurs pénalisés le long du spectre : en cherchant un facteur multiplicatif par bloc spectral de taille suffisamment grande (de l'ordre de la centaine de feuillets), on peut s'assurer que le flux moyen (sur le bloc) est conservé, tout en conservant le bénéfice de la régularisation. Ce facteur de correction est obtenu de la manière suivante : on cherche une matrice diagonale $\boldsymbol{F} = \text{diag}(\{f_j\}_{1\leq j\leq k})$ telle qu'on ait pour le continuum $\boldsymbol{Y}^c = \widetilde{\boldsymbol{U}}\boldsymbol{F}\widehat{\boldsymbol{D}}^c$. La matrice $\boldsymbol{F}$ étant diagonale cela revient à définir

$$f_j = \frac{\left\langle \left(\boldsymbol{Y}^c \left(\widehat{\boldsymbol{D}}^c\right)^+\right)_j, \widetilde{\boldsymbol{U}}_j \right\rangle}{||\widetilde{\boldsymbol{U}}_j||_2^2}, \text{ pour } 1 \leq j \leq k. \tag{3.34}$$

Notons que si cette correction, qui se fait au sens des moindres carrés, était appliquée feuillet par feuillet, elle annulerait tout simplement l'effet de la régularisation ridge.

L'approche régularisée d'estimation des spectres pour une image HST i au sein d'un bloc spectral j à FSF constante est résumée dans l'algorithme 4.

Remarque 7
Afin de prendre en compte les résidus de la soustraction du spectre de l'atmosphère, on ajoute tout au long de la procédure de démélange, l'estimation d'un spectre constant spatialement dans la zone considérée, correspondant au spectre du fond de ciel. Ce spectre peut être vu comme l'intercept du problème de régression étudié.

3.3.6 Résultats sur données simulées

Sur les figures 3.15 et 3.16 sont montrés, à partir de données simulées, les différents spectres intermédiaires produits tant au niveau pixelique (figure 3.15) qu'au niveau des sources (figure 3.16). La figure 3.15 illustre les différents étapes de séparation entre raie et continuum subies par un spectre (ici situé au centre de la source 1). On commence par estimer le continuum (figure 3.15b), par exemple par un filtre médian, dont la taille de la fenêtre est suffisamment grande (e.g. trois fois plus grande) par rapport aux largeurs spectrales des raies, puis on en déduit par soustraction le spectre des raies pour ce pixel(figure 3.15c). On cherche ensuite à estimer les raies des spectres des objets. Pour cela on cherche à détecter tous les supports de raies potentielles. Comme décrit dans l'algorithme 2, afin de "robustifier" cette détection et limiter le temps de calcul, on effectue cette détection sur un ensemble de spectres représentatifs de l'ensemble de données : à partir des spectres de raies des pixels, on construit des estimations des spectres de chaque objet par une somme pondérée par la carte d'intensité de l'objet. On

Algorithme 4 Procédure régularisée d'estimation des spectres

Entrée : matrice de données $\boldsymbol{Y} \leftarrow \boldsymbol{Y}_j$, matrice d'intensité $\widetilde{\boldsymbol{U}} \leftarrow \widetilde{\boldsymbol{U}}_{i,j}$

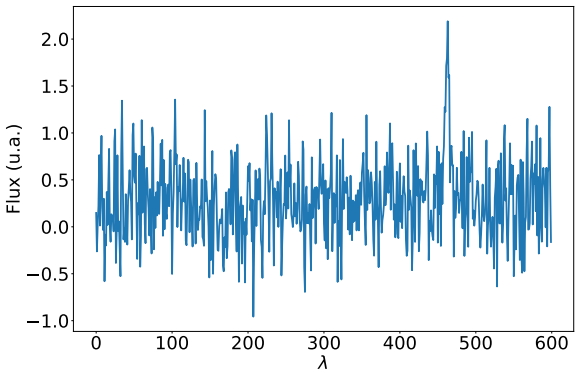
Estimation de $\boldsymbol{Y}^c$ ▷ Estimation du continuum par filtrage robuste
Calcul des raies $\boldsymbol{Y}^r = \boldsymbol{Y} - \boldsymbol{Y}^c$ ▷ Soustraction du continuum
Calcul de $\mathcal{E}$ à l'aide de l'Alg. 2 ▷ Détection des raies
for raie l **do**
    Calcul du modèle $\mathcal{M}_l$ à l'aide de l'eq. (3.27) ▷ Sélection du modèle
    Calcul de $\widehat{\boldsymbol{D}}_l^r$ par l'équation (3.28)
Construction de $\widehat{\boldsymbol{D}}^r$ ▷ Concaténation spectrale
Calcul de $\boldsymbol{Y}^c = \boldsymbol{Y} - \boldsymbol{Y}^r$
for bloc spectral b **do**
    Calcul de α_{m+1s} à l'aide de l'alg. 3 ▷ Estimation GCV du paramètre de régul.
    Calcul de $\widehat{\boldsymbol{D}}^r$ par la formule (3.29) avec $\alpha = \alpha_{m+1s}$ ▷ Estimation continuum par régul. ridge
Construction de $\widehat{\boldsymbol{D}}^c$ ▷ Concaténation spectrale
Calcul de $\boldsymbol{F}$ à l'aide de l'équation (3.34) ▷ Facteurs de correction de flux
Calcul de $\widehat{\boldsymbol{D}} = \widehat{\boldsymbol{D}}^r + \boldsymbol{F}\widehat{\boldsymbol{D}}^c$ ▷ Combinaison des raies et du continuum

Sortie : $\widehat{\boldsymbol{D}}$ ▷ Spectres des sources (pour l'image HST i et la bande spectrale j à FSF constante)

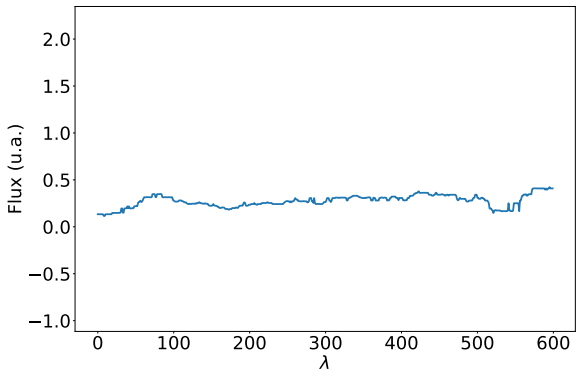
détecte ensuite les raies potentielles (figure 3.16a) sur ces spectres représentatifs. Sur chacun des supports détectés, on effectue alors une régression avec régularisation BIC décrite précédemment et on obtient les spectres de raies de chacun des objets (figure 3.16b). On peut voir que pour chaque support, seul un des deux spectres est non-nul, grâce à la régularisation BIC. Notons également que du fait de la soustraction imparfaite du continuum, la deuxième raie (en bleue) comporte un léger artefact négatif sur un bord. Cela sera compensé lors de l'estimation du continuum ci-après (puisqu'on cherche ensuite à estimer le tout le signal restant). A partir de cette estimation, on déduit, pour chaque pixel, le continuum restant à expliquer (figure 3.15c). On applique ensuite la régression avec régularisation ridge sur ces continuums et on obtient les spectres de continuum des objets (figure 3.16c). En combinant les raies et le continuum, on obtient les spectres finaux des deux sources (figure 3.16d).

Si on analyse désormais l'apport de ces régularisations sur les performances, on peut voir sur la figure 3.17 que l'ajout de la régularisation permet de fortement améliorer la robustesse des estimations lorsque le conditionnement se dégrade : les spectres estimés possèdent moins d'artefacts anticorrélés (figures 3.17a et 3.17b) et sont plus proches de la vérité-terrain (figure 3.17c).

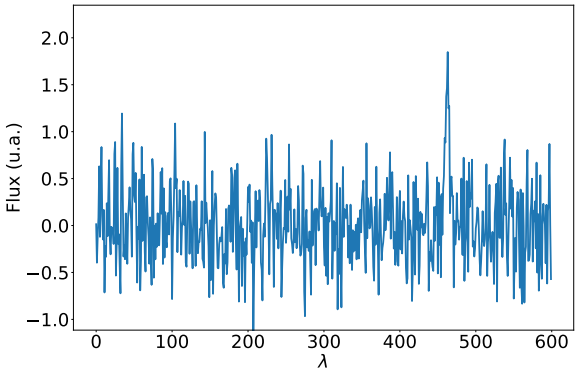
Notons que cette méthode reste non-supervisée puisque la régularisation se fait sans paramètre pour les raies et avec un paramètre estimé par validation croisée pour le continuum.



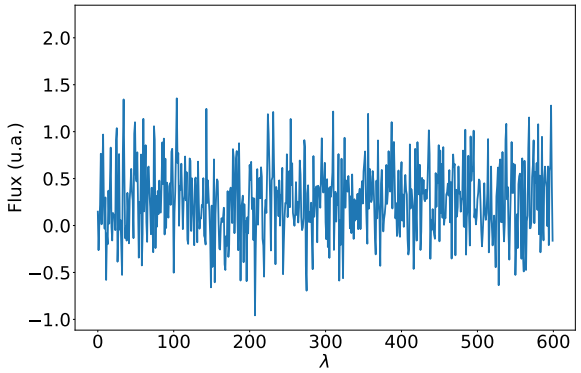
(a) Spectre initial d'un pixel au centre de la source 1



(b) Estimation du continuum de ce pixel

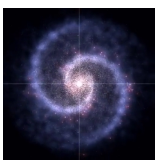


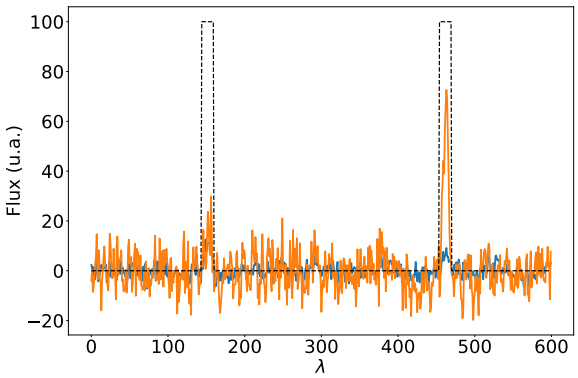
(c) Estimation des raies de ce pixel (spectre initial - continuum)



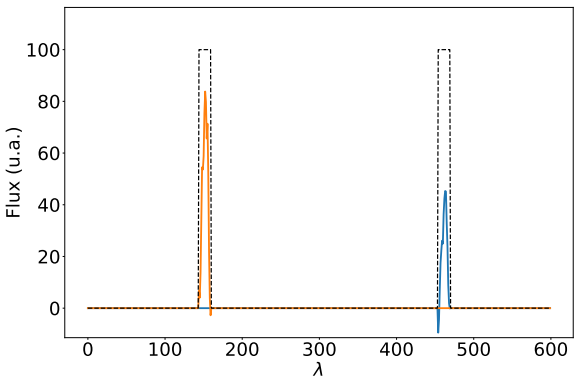
(d) Continuum restant pour un pixel (spectre initial - raie estimée

FIGURE 3.15 – Séparation des raies et du continuum pour un spectre/pixel situé au centre de la source 1.

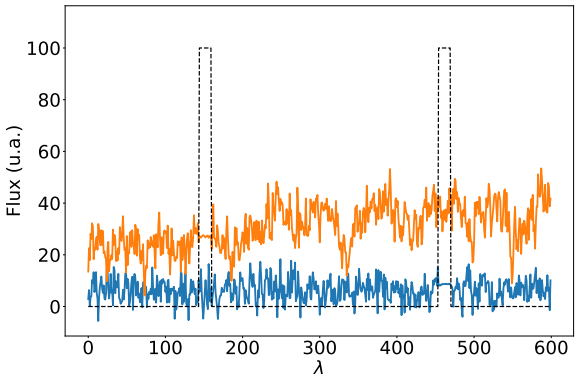




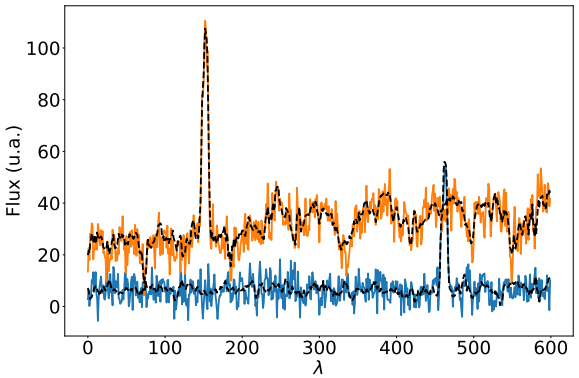
(a) Détection des raies sur des estimations simples des spectres des objets. En pointillé noir les supports des raies détectées.



(b) Estimation des raies des spectres des objets. En pointillés noirs les supports des raies détectées.

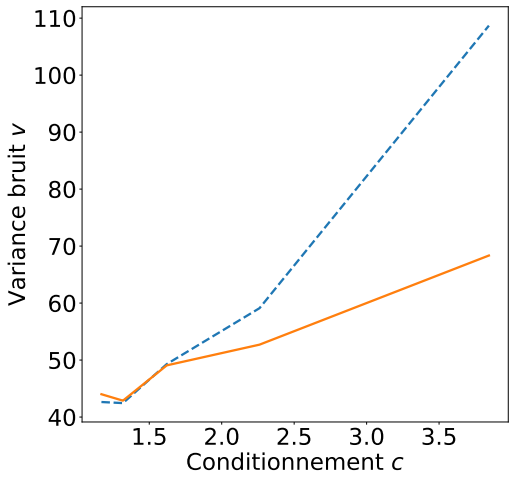


(c) Estimation du continuum des spectres des objets.

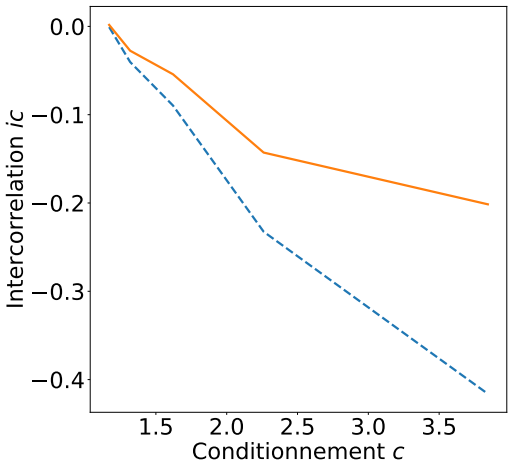


(d) Estimation complète des spectres des objets. En pointillés noirs la vérité-terrain.

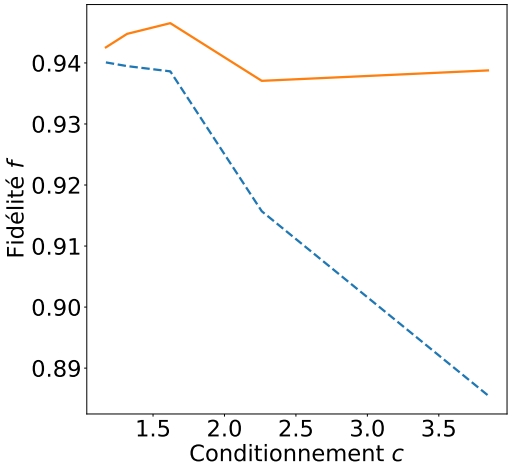
FIGURE 3.16 – Estimation des spectres des objets



(a) Evolution de la variance du bruit en fonction du conditionnement

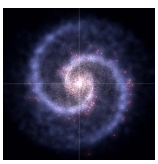


(b) Evolution de l'intercorrélation entre les spectres des objets en fonction du conditionnement



(c) Evolution de la fidélité à la vérité-terrain en fonction du conditionnement

FIGURE 3.17 – Performances en fonction du conditionnement. Les pointillés bleu indiquent l'absence de régularisation ; les traits pleins oranges indiquent l'utilisation de la régularisation. Résultats moyennés sur 20 simulations Monte-Carlo.



Résumé

Dans ce chapitre nous avons proposé une méthode de démélange spectral des sources MUSE à l'aide des données HST. Dans un premier temps, nous avons mis en place une méthode de régression linéaire non régularisée, s'appuyant sur des cartes d'abondance construites grâce aux images HST. Cette première approche a été testée sur données simulées, ce qui a permis de mettre en exergue une faiblesse importante : le problème de démélange devient rapidement mal posé lorsque que les sources à démélanger sont proches spatialement. Face aux limitations de cette première approche, un ensemble de régularisations adéquates a été mis en place et validé sur données simulées. La stratégie choisie consiste à traiter de façon séparée le continuum spectral, dont l'estimation est régularisée par une pénalité de type ridge, et les raies spectrales, qui sont obtenues à l'aide d'une régularisation par parcimonie. Dans le chapitre suivant, nous allons désormais appliquer ces méthodes sur les données MUSE du champ profond UDF.

Chapitre 4

Application sur données réelles

Sommaire

Dans ce chapitre, la méthode de démélange développée dans le chapitre précédent est testée sur quelques situations de mélange dans le champ profond UDF observé par MUSE.

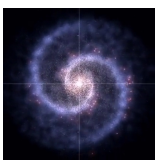
4.1 Données

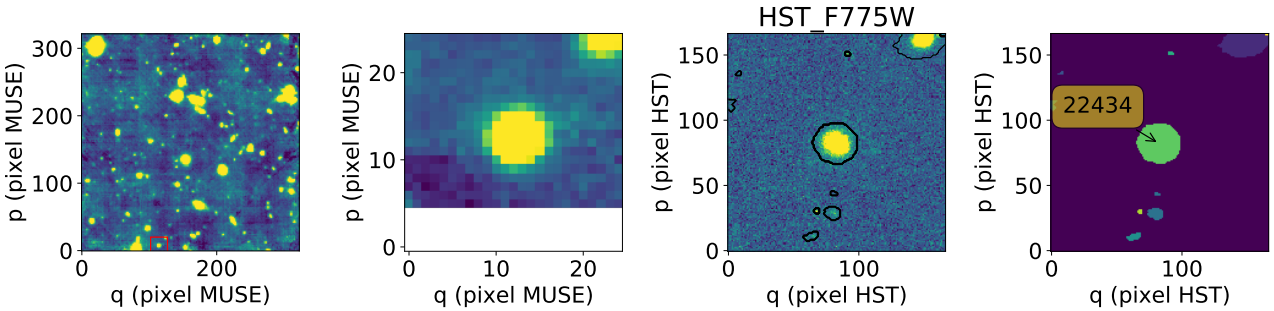
On montre ici les résultats de la méthode de démélange pour quelques objets présents dans le champ central de l'UDF observé par MUSE. Un catalogue d'objets a été construit sur les données MUSE, en s'appuyant sur le catalogue HST et en explorant manuellement les données MUSE. Les objets étant référencés à la fois par un catalogue HST et un catalogue MUSE, on note IDM l'identifiant de l'objet dans le catalogue MUSE et IDH son pendant dans le catalogue HST. Pour plus de détails sur la création de ce catalogue, le lecteur peut se référer à [Bacon et al. 2017]. Les objets sélectionnés ici ont été "démélangés" manuellement par des experts astrophysiciens en combinant une étude des images en bande-étroite (somme uniquement sur les longueurs d'onde d'une raie) autour de chaque raie et une étude des combinaisons possibles de raies pour chaque objet, à l'aide de modèles de spectres. **Cela permet toutefois uniquement d'associer les raies à des objets et non de démélanger le continuum pour lequel il n'y a donc pas de vérité-terrain.**

La méthode décrite dans le chapitre précédent est appliquée sur les objets de façon non supervisée. Les seuls paramètres fixés concernent la détection des supports de raies et ont été fixé après un apprentissage sur les données (cf paragraphe 3.3.4.1). Par soucis de concision, la méthode proposée est par la suite référencée comme ODHIN (pour Optimal Deblending of Hyperspectral ImagiNg). Elle est comparée au spectre servant jusqu'alors de référence pour les astronomes. Ce spectre est estimé pour chaque objet du catalogue MUSE de la manière suivante : on obtient un masque binaire en convoluant la région de l'objet HST correspondant sur la carte de segmentation du HST avec la FSF de MUSE. On estime alors le spectre de l'objet en sommant les spectres de tous les pixels à l'intérieur de ce masque. Ce spectre est référencé par la suite par le nom 'MUSE_TOT_SKYSUB'.

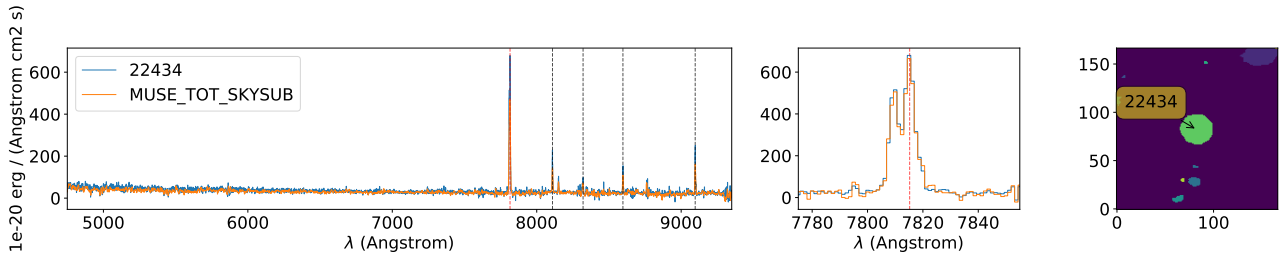
4.2 Résultats

On peut voir tout d'abord sur la figure 4.1 qu'en l'absence de situation de mélange, la méthode proposée estime une solution identique à l'approche classiquement utilisée par les astronomes. On s'assure ainsi notamment que le flux spectral est bien conservé.



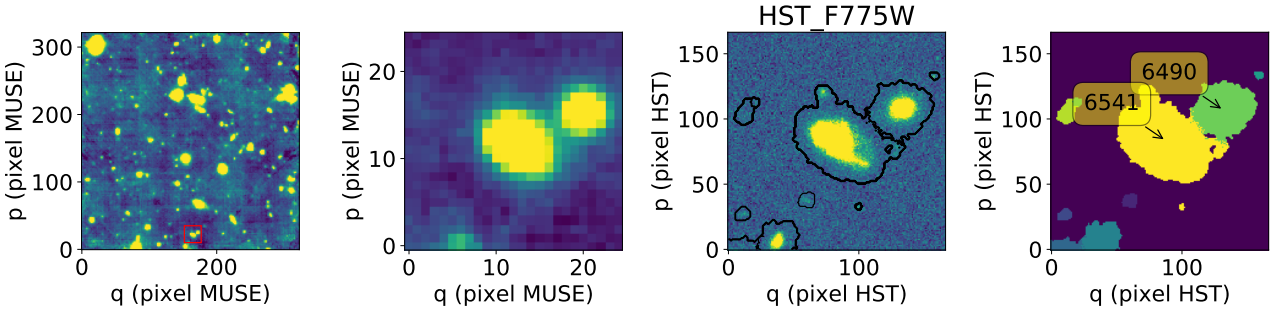


(a) Images MUSE et HST de l'objet IDM30. De gauche à droite : position de l'objet dans le champ MUSE UDF-10, image blanche MUSE, image HST (filtre 775W) et carte de segmentation.

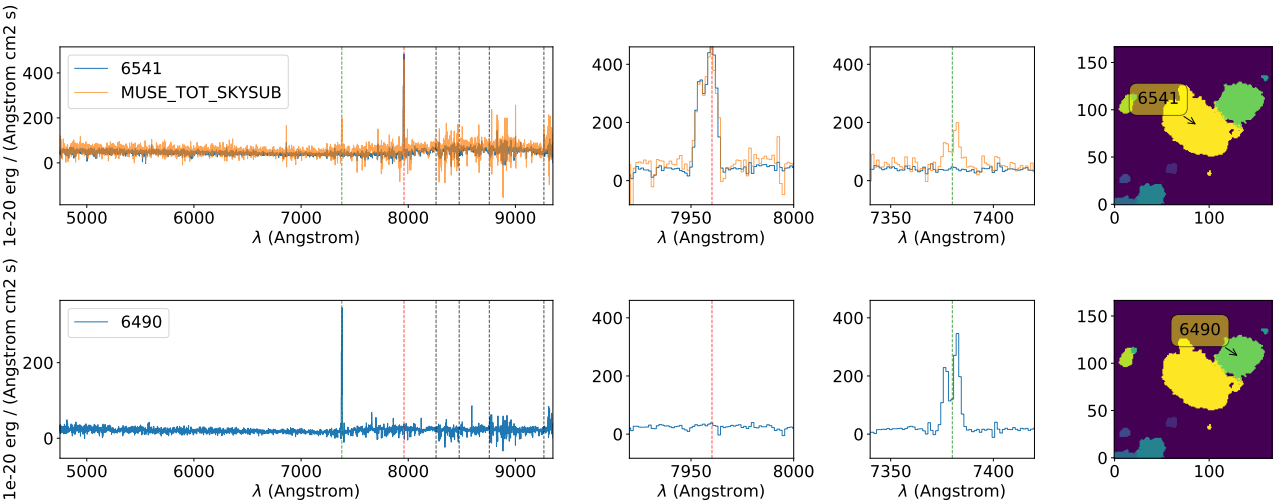


(b) Spectre estimé IDM30. De gauche à droite : spectre estimé par ODHIN (bleu) et spectre 'MUSE_TOT_SKYSUB' (orange), zoom autour d'une raie, carte de segmentation

FIGURE 4.1 – Estimation du spectre de l'objet 30 du catalogue MUSE et 22434 de HST (IDM30-HST22434). Sur la première ligne : images MUSE et HST de l'objet. Sur la seconde ligne : comparaison des spectres estimés.

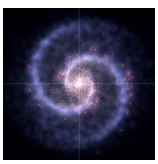


(a) Images MUSE et HST de l'objet IDM26. De gauche à droite : position de l'objet dans le champ MUSE UDF-10, image blanche MUSE, image HST (filtre 775W) et carte de segmentation.



(b) Spectres estimés. Première ligne, de gauche à droite : spectre estimé par ODHIN pour IDH6541 (bleu) et spectre 'MUSE_TOT_SKYSUB' (orange) ; zoom autour d'une raie associée à IDH6541 (raie en rouge) ; zoom autour d'une raie associée à IDH6490 (raie en vert) ; carte de segmentation. Seconde ligne : spectre estimé par ODHIN pour IDH6490 (bleu) ; zoom autour d'une raie associée à IDH6541 (raie en rouge) ; zoom autour d'une raie associée à IDH6490 (raie en vert) ; carte de segmentation.

FIGURE 4.2 – Estimation du spectre de l'objet 26 du catalogue MUSE, correspondant aux objets 6541 et 6490 du HST. (a) : images MUSE et HST de l'objet. (b) : comparaison des spectres estimés. Les raies appartenant à IDH6541 sont en noir (ainsi que la raie rouge étudiée), la raie étudiée appartenant à IDH6490 est en vert.



Sur la figure 4.2, on peut voir un example de démélange fonctionnel autour de l'objet IDM26, correspondant aux objets IDH6541 et IDH6490 du catalogue HST. On peut voir notamment que la contamination de la raie autour de 7400Å, très présente sur le spectre 'MUSE_TOT_SKYSUB', disparaît complètement du spectre de IDH6541 estimé par ODHIN. La figure 4.3 présente les résultats de reconstruction après démélange autour de l'objet IDM26. On peut voir que la reconstruction engendre des résidus spatiaux faibles (similaires au niveau de bruit) (figure 4.3.a) et que les deux raies étudiées sont bien associées au bon objet (figure 4.3.b).

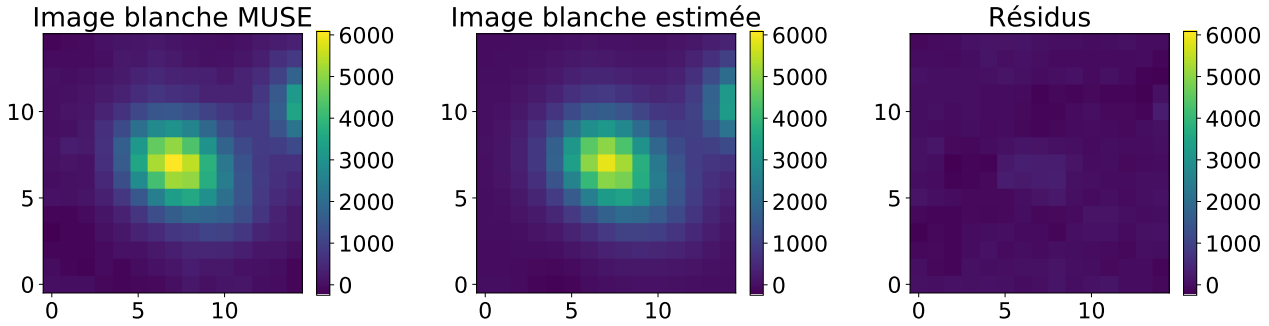
On peut voir sur la figure 4.4 un second example de démélange fonctionnel autour de l'objet IDM69, correspondant aux objets IDH22350 et IDH22351 du catalogue HST. On peut voir notamment que la raie de l'objet IDH22350 autour de 7800Å est bien associée au bon spectre. On peut également noter que le flux du continuum est réparti entre les deux objets (baisse du flux associé à l'objet IDH22351).

On peut voir sur la figure 4.5 un exemple de démélange où la méthode ne réussit qu'imparfaitement. La contamination de l'objet IDH23794 sur l'objet IDH60001 est fortement réduite par rapport au résultat obtenu sur le spectre 'MUSE_TOT_SKYSUB' (à la fois au niveau des raies et du continuum) mais toujours présente. Il est à noter que la source IDH60001 n'appartient pas au catalogue HST de [Rafelski et al. 2015] mais a été détectée grâce aux données MUSE, la segmentation sur HST étant alors faite suite à cette découverte à l'aide du logiciel NoiseChisel (Akhlaghi et Ichikawa 2015).

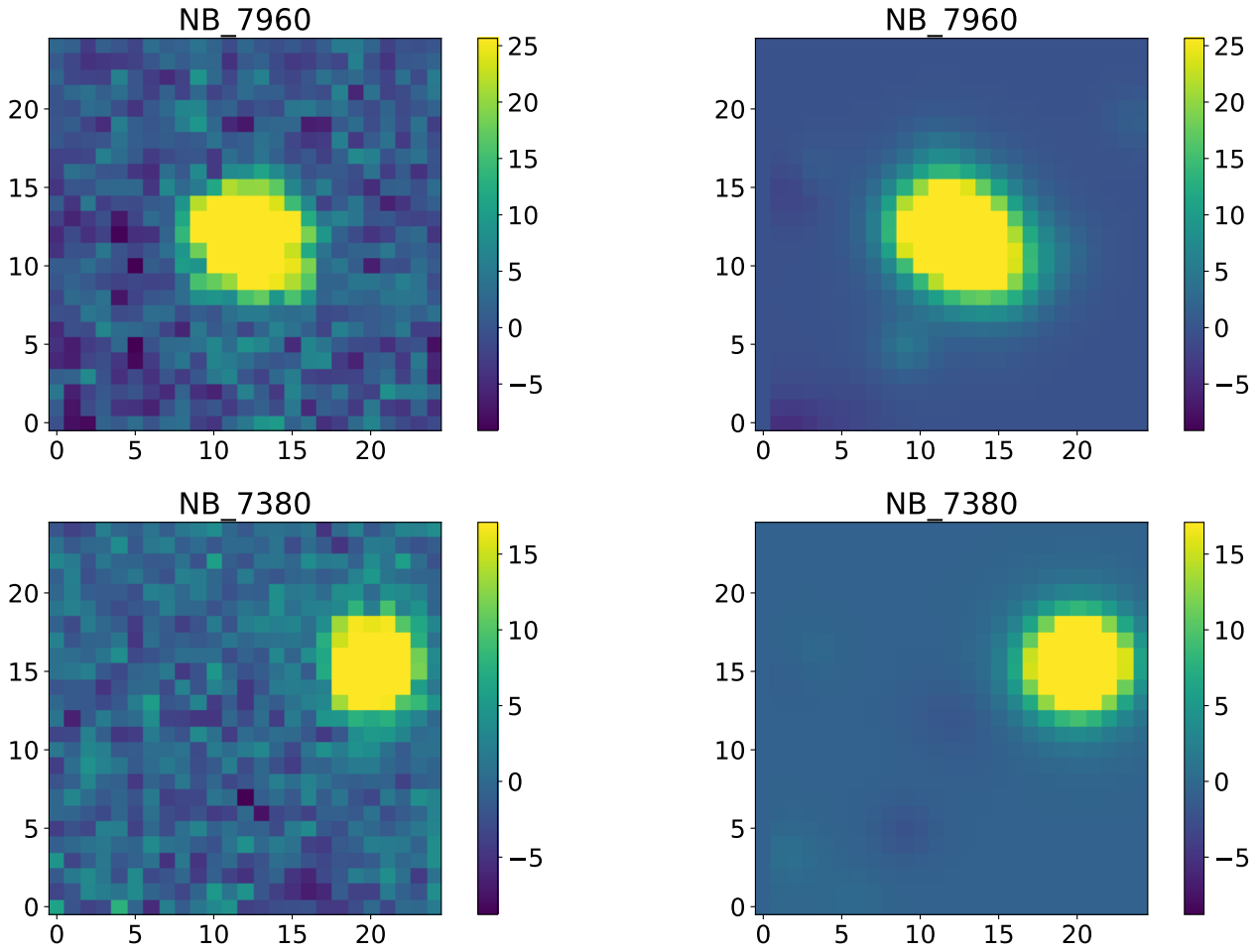
Ces quelques exemples semblent montrer que la méthode fonctionne bien sur les données réelles. Elle a donc été transmise aux astronomes du CRAL et va désormais être utilisée sur l'ensemble des sources de la mosaïque UDF. A l'heure actuelle, sur une centaine de sources testées (possédant toutes une certaine vérité terrain issue de l'analyse manuelle par les experts), environ 70% sont considérées comme succès, les 30% restants ayant des résultats plus nuancés, notamment dès lors que l'information du continuum obtenue sur HST est insuffisante pour traduire avec précision le comportement spectral de la source. L'analyse de ces résultats et leurs interprétations astrophysiques seront développés dans un article de journal, actuellement en préparation, à destination de la communauté astrophysique.

Implémentation et coût calculatoire

Le code a été développé en Python, en s'appuyant sur les bibliothèques *numpy* et *scipy*, ainsi que sur la bibliothèque *mpdaf* du consortium MUSE. Actuellement, le traitement sans régularisation d'une région 25×25 pixels (à la résolution MUSE) avec une quinzaine de sources et 10 blocs spectraux prend de l'ordre de 40s sur un processeur Intel à 8 coeurs cadencés à 3GHz. Le traitement avec régularisation prend de l'ordre de 1min pour une région 25 par 25 pixels. Un tiers est dû à la transformation des cartes d'intensité de HST à MUSE. Ces opérations sont basées sur des transformées en Fourier rapides et évoluent donc en $O(bN\log(N)$ avec N le nombre de pixels HST et b le nombre de blocs spectraux à FSF constante. Le restant est dû à l'inversion de systèmes de taille $n \times k$ où k est le nombre de sources et n le nombre de pixels MUSE. En l'absence de régularisation, la complexité de l'inversion au sens des moindres carrés implémentée dans le paquet python *numpy* est en $k^2 \times n$. L'ajout de la régularisation entraîne un surcoût calculatoire qui devient non négligeable lorsque le nombre de sources augmente. La régularisation BIC des raies, consiste à faire autant d'inversions moindres carrés du système que de sources présentes dans la zone et reste donc linéaire en n ; la régularisation ridge du

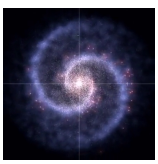


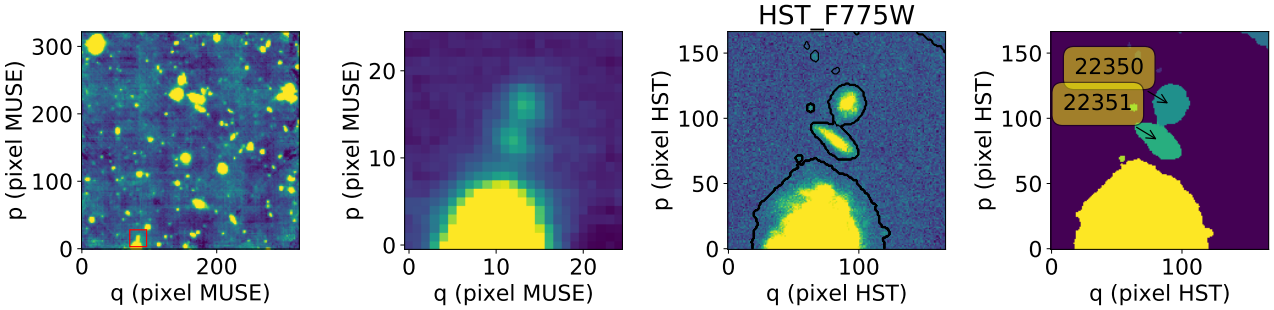
(a) Résidus. De gauche à droite : image blanche MUSE, image blanche reconstruite et résidus (sommés le long des bandes spectrales).



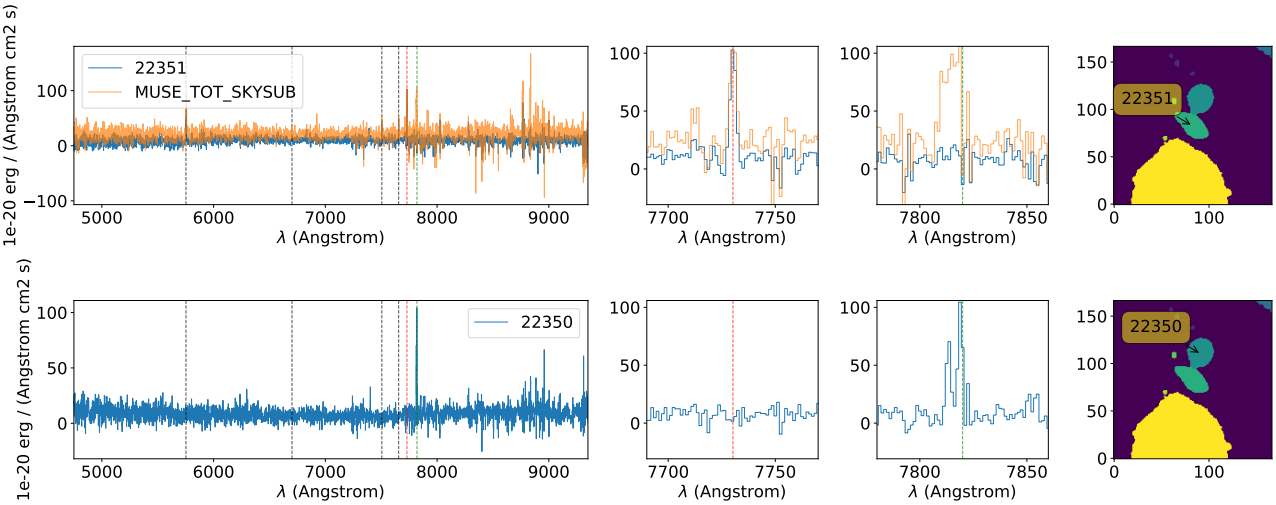
(b) Images bandes étroites (±10 feuillets). Première colonne : image MUSE ; seconde colonne : image estimée par ODHIN. Première ligne : bande étroite autour de 7960Å ; seconde ligne : bande étroite autour de 7380Å.

FIGURE 4.3 – Estimation de la réponse spatiale de l'objet IDM26, correspondant aux objets IDH6541 et IDH6490 du catalogue HST. (a) : comparaisons des images blanches. (b) : comparaisons des images en bande étroite.



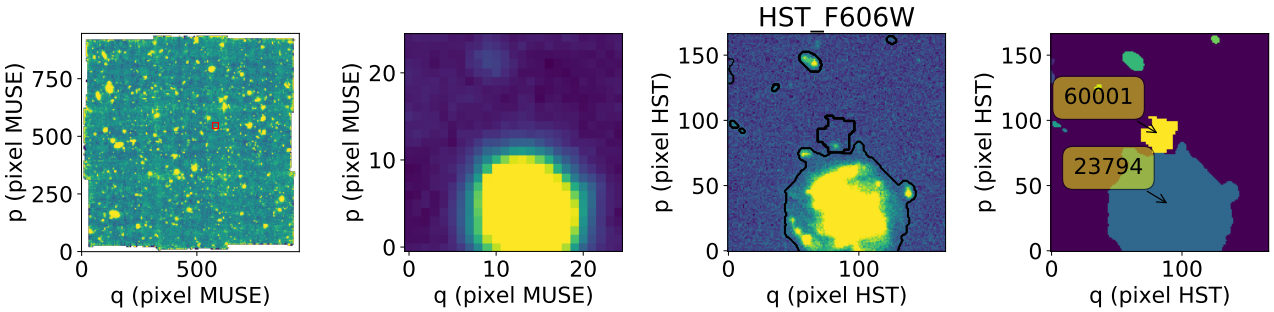


(a) Images MUSE et HST de l'objet IDM69. De gauche à droite : position de l'objet dans le champ MUSE UDF-10, image blanche MUSE, image HST (filtre 775W) et carte de segmentation.

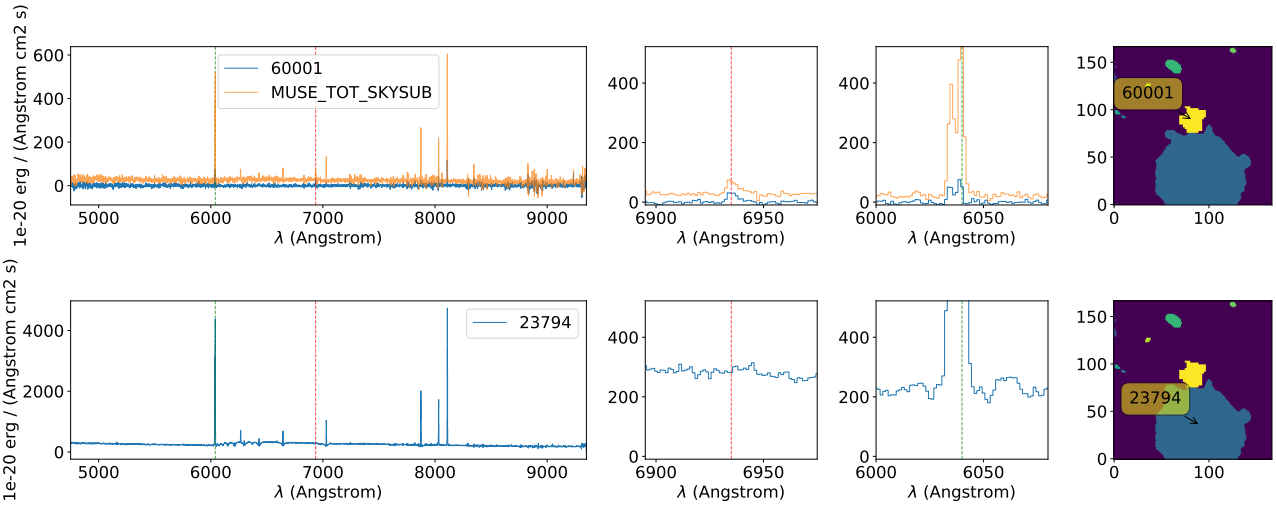


(b) Spectres estimés. Première ligne, de gauche à droite : spectre estimé par ODHIN pour IDH22351 (bleu) et spectre 'MUSE_TOT_SKYSUB' (orange) ; zoom autour d'une raie associée à IDH22351 (raie en rouge) ; zoom autour d'une raie associée à IDH22350 (raie en vert) ; carte de segmentation. Seconde ligne : spectre estimé par ODHIN pour IDH22350 (bleu) ; zoom autour d'une raie associée à IDH22351 (raie en rouge) ; zoom autour d'une raie associée à IDH22350 (raie en vert) ; carte de segmentation.

FIGURE 4.4 – Estimation du spectre de l'objet 69 du catalogue MUSE, correspondant aux objets 22350 et 22351 de HST. (a) : images MUSE et HST de l'objet. (b) : comparaison des spectres estimés. Les raies appartenant à IDH22351 sont en noir (ainsi que la raie rouge étudiée), la raie étudiée appartenant à IDH22350 est en vert.

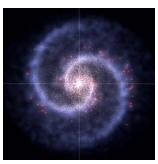


(a) Images MUSE et HST de l'objet IDM6313. De gauche à droite : position de l'objet dans le champ MUSE UDF-10, image blanche MUSE, image HST (filtre 775W) et carte de segmentation.



(b) Spectres estimés. Première ligne, de gauche à droite : spectre estimé par ODHIN pour IDH60001 (bleu) et spectre 'MUSE_TOT_SKYSUB' (orange) ; zoom autour d'une raie associée à IDH6541 (raie en rouge) ; zoom autour d'une raie associée à IDH23794 (raie en vert) ; carte de segmentation. Seconde ligne : spectre estimé par ODHIN pour IDH23794 (bleu) ; zoom autour d'une raie associée à IDH60001 (raie en rouge) ; zoom autour d'une raie associée à IDH23794 (raie en vert) ; carte de segmentation.

FIGURE 4.5 – Estimation du spectre de l'objet 6313 du catalogue MUSE, correspondant aux objets 60001 et 23794 de HST. (a) : images MUSE et HST de l'objet. (b) : comparaison des spectres estimés.



continuum a le même coût que l'inversion moindres carrés.

Plus précisément, la régularisation ridge a un coût en $a\lambda k^2 n$ avec λ le nombre total de feuillets, et a le nombre de paramètres α à tester (en pratique fixé à 50) et la régularisation BIC peut être effectué en $Rk^2 n$, où R est le nombre de raies détectées, si implémenté de façon optimale sous forme récursive (l'implémentation actuelle est en $Rk^3 n$). L'objectif est à terme de pouvoir traiter tout un champ MUSE voir une mosaïque de champ, avant toute détection et définition de zone d'intérêt par source. La méthode proposée, bien que peu calculatoire passera déjà tout juste à l'échelle et nécessitera vraisemblablement une découpe du champ en un certain nombre de sous-champs pour rester exécutable en quelques heures. Le temps d'exécution avec régularisation peut être vérifié en pratique sur <https ://phd.rbacher.fr/>.

Bilan et perspectives de la partie I

Bilan

Dans cette partie, nous nous sommes intéressés à la problématique du démélange spectral de sources galactiques présentes dans les champs profonds observés par MUSE.

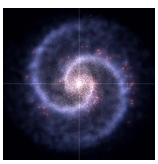
Dans un premier temps, nous avons mis en place une méthode de régression linéaire non régularisée, s'appuyant sur des cartes d'abondance construites grâce aux images HST. Cette première approche a été testée sur données simulées, ce qui a permis de mettre en exergue une faiblesse importante : le problème de démélange devient rapidement mal posé lorsque que les sources à démélanger sont proches spatialement.

Face aux limitations de cette première approche, un ensemble de régularisations adéquates a été mis en place et validé sur données simulées. **La stratégie choisie consiste à traiter de façon séparée le continuum spectral, dont l'estimation est régularisée par une pénalité de type ridge, et les raies spectrales, qui sont obtenues à l'aide d'une régularisation par parcimonie.** Cette approche régularisée a alors été testée avec succès sur un échantillon de sources ayant été analysées manuellement par des experts. Notons que cette stratégie de régularisation a notamment permis d'obtenir une approche suffisamment robuste sur données réelles, là où des tests effectués d'une régularisation ridge par validation croisée, appliquée globalement sur l'ensemble raies plus continuum des spectres, ne parvenaient pas assurer des résultats suffisamment consistants d'un objet à l'autre.

La méthode proposée a été implémentée de façon à s'intégrer dans la bibliothèque logicielle développée au sein du consortium MUSE. Cette implémentation a alors été testée sur les données réelles de l'observation MUSE de l'UDF. Les résultats obtenus sur données réelles sont probants puisqu'on reconstruit pour la plupart des sources testées des spectres en accord avec l'analyse des experts. La méthode rencontre toutefois ses limites dans un certain nombre de cas, principalement lorsque les images HST (qui contiennent l'information du continuum spectral) ne reflètent pas avec précision la distribution spatiale des raies. **Une limitation importante de la méthode développée est le postulat d'invariabilité spectrale au sein d'une source.** En effet ce modèle ne tient plus dans plusieurs cas : certaines sources brillantes étendues possèdent un champ de vitesse donc une variabilité (effet Doppler) et certaines sources lointaines très étendues peuvent également posséder une variabilité spectrale. Il s'agit d'une variabilité qui se traduit par une déformation fortement non-linéaire, une dilatation du spectre, qu'il est donc difficile d'intégrer dans le modèle linéaire actuel. La présente étude se limite ainsi à obtenir pour chaque objet une signature spectrale "moyenne". Par rapport à des méthodes de déconvolution plus générales, ou à des approches de type pansharpening, la méthode proposée s'attaque à un problème mieux posé du fait des informations *a priori* venant de HST. L'inconvénient en contre-partie est que la méthode est fortement contrainte et donc limitée à la qualité de cette information *a priori*.

Perspectives

De nombreuses perspectives sont envisageables suite à ces travaux, certaines ayant d'ailleurs déjà été abordées de façon exploratoire pendant la thèse.



Extension 1 : ajout d'une régularisation de non-négativité douce

Problématique

Du fait des prétraitements, notamment pour soustraire la contribution de l'atmosphère, les données obtenues ne sont pas toujours positives. Appliquer directement une contrainte de non-négativité sur les spectres estimés entraînerait donc un biais dans le flux estimé (suppression des composantes bruitées négatives et conservation des composantes bruitées positives). Nous voulons toutefois éviter des pics fortement négatifs qui n'ont a priori pas de sens physique.

Méthode envisagée

On peut envisager d'imposer cette régularisation de la manière suivante :

- Pénalisation des fortes valeurs négatives sous la forme d'un terme de pénalité $f(x) = (x - |x|)^2$.
- Implémentation par ADMM de la résolution de $\min_{\boldsymbol{D}_{i,l}} ||\tilde{\boldsymbol{Y}}_l - \boldsymbol{D}_{i,l}\tilde{\boldsymbol{U}}_{i,l}||_2^2 + \alpha||f(\boldsymbol{D}_{i,l})||_2$
- Choix du paramètre de régularisation par GCV

Extension 2 : ajout d'informations physiques

Problématique

A l'heure actuelle, une fois les spectres estimés, la validation par les experts se fait en cherchant à estimer les redshifts des objets. Cette estimation se fait à l'aide de modèles (ou *templates*) de combinaison de raies d'absorption et d'émission et consiste à trouver conjointement le template le plus adapté et la valeur de redshift associée. Cela permet notamment d'imposer des contraintes fortes sur la compatibilité de certaines raies au sein d'un même spectre. Il est donc pertinent de chercher à introduire ces contraintes dans le processus de démélange afin de favoriser les solutions les plus vraisemblables physiquement.

Méthode envisagée

Une piste envisagée consiste alors à ajouter un terme de pénalité imposant une parcimonie des spectres estimées sur un dictionnaire composé des templates dupliqués à tous les redshifts possibles. Cela pose plusieurs difficultés auquel il va falloir répondre :

- Les templates ne prenant pas en compte le continu, la décomposition des spectres sur le dictionnaire doit se faire après une étape de soustraction du continu du spectre.
- Le nombre d'atomes dans le dictionnaire peut devenir assez important (pour assurer un partitionnement assez fin en redshift il faudrait de l'ordre de 20000 atomes par template).

Une approche itérative peut être envisagée pour surmonter ces difficultés :

- Premier démélange sans a priori
- Construction de templates spectraux de combinaison de raies
- Sélection des translatées de templates corrélées aux spectres estimés pour construire un dictionnaire
- Nouveau démélange avec contraintes spectrales du dictionnaire

Extension 3 : Démélange en utilisant conjointement MUSE et HST

Problématique

La méthode actuelle n'exploite que le pouvoir de séparation spatial du HST pour effectuer le démélange. On peut toutefois chercher à exploiter directement l'information spectrale en exploitant une hypothèse de décomposition parcimonieuse du cube super-résolu $\boldsymbol{X}$ sur un dictionnaire de spectres. Notons que cela peut permettre également de relâcher la contrainte sur la matrice de mélange $\boldsymbol{U}$, considérée jusqu'à présent comme parfaitement déterminée par HST.

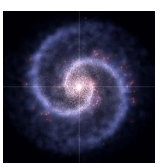
Méthode envisagée

On va chercher à obtenir conjointement le cube super-résolu $\boldsymbol{X}$ et une représentation parcimonieuse de $\boldsymbol{X}$ en un dictionnaire $\boldsymbol{D}$ et une matrice de mélange $\boldsymbol{U}$. Dans un premier temps on peut initialiser $\boldsymbol{U}$ et $\boldsymbol{D}$ via la méthode de démélange décrite précédemment. On peut ensuite estimer $\boldsymbol{X}$ en tant que solution de :

$$\min_{\boldsymbol{X}} \underbrace{||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}||_2^2}_{Data} + \alpha \underbrace{||\boldsymbol{X} - \boldsymbol{D}\boldsymbol{U}||_2^2}_{Reg.} \tag{4.1}$$

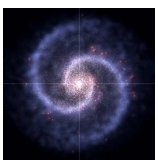
où α est un paramètre de régularisation.

Une fois obtenu un cube super-résolu, on peut chercher à obtenir une nouvelle représentation parcimonieuse des spectres via une méthode d'apprentissage comme K-SVD Aharon et al. 2006. L'utilisation de K-SVD permet d'imposer aisément la contrainte sur les supports spatiaux (lors de l'étape de reconstruction par Orthogonal Matching Pursuit). De plus cela permet également d'avoir un contrôle direct sur la parcimonie de la solution. On peut en fait itérer le processus en alternant entre des étapes d'apprentissage de ($\boldsymbol{D}$,$\boldsymbol{U}$) sur $\boldsymbol{X}$ et des étapes d'estimation de $\boldsymbol{X}$.



Deuxième partie

Détection de sources étendues



Notations et équations - partie II

Notations

- $\mathcal{H}_0$: hypothèse nulle (pas de signal)
- $\mathcal{H}_1$: hypothèse alternative (signal d'intérêt)
- $\boldsymbol{y}$: vecteur d'observation
- $\boldsymbol{\epsilon}$: vecteur bruit
- $\boldsymbol{x}$: signal d'intérêt
- $\boldsymbol{D}$: dictionnaire composé d'atomes $\boldsymbol{d}_j$
- $S()$: mesure de similarité
- F : fonction de répartition
- G : fonction de survie
- $\overline{F}$: fonction de répartition empirique
- $\overline{G}$: fonction de survie empirique
- π_0 : proportion d'échantillons nuls
- n : nombre d'échantillons/pixels
- η : niveau de seuil
- $a \vee b$: $\max(a, b)$
- $\text{sgn}(a)$: signe de a

Equations

$$\begin{cases} \mathcal{H}_0 : \boldsymbol{y} = \boldsymbol{\epsilon}, \\ \mathcal{H}_1 : \boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\epsilon}, \end{cases} \tag{6.1}$$

$$\begin{aligned} &\boldsymbol{x} \approx a_{i_1}\boldsymbol{d}_{i_1} + \ldots + a_{i_k}\boldsymbol{d}_{i_k}, \\ &\text{avec } a_{i_j} > 0, \quad \text{pour } 1 \leq j \leq k, \end{aligned} \tag{6.2}$$

$$\begin{cases} \mathcal{H}_0 : a_1 = a_2 = \ldots = a_m = 0, \\ \mathcal{H}_1 : \text{ au moins un } a_i > 0, \end{cases} \tag{6.3}$$

$$S_{FA}(\boldsymbol{y}, \boldsymbol{d}) \equiv \left\langle \frac{\boldsymbol{d}}{||\boldsymbol{d}||}, \boldsymbol{y} \right\rangle = \boldsymbol{d}^T\boldsymbol{y} \tag{6.4}$$

$$S_{SAD}(\boldsymbol{y}, \boldsymbol{d}) \equiv \frac{\langle \boldsymbol{d}, \boldsymbol{y} \rangle}{||\boldsymbol{d}||.||\boldsymbol{y}||} = \frac{\boldsymbol{d}^T\boldsymbol{y}}{||\boldsymbol{y}||} \tag{6.5}$$

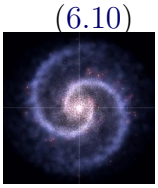
$$T_{\max}(\boldsymbol{y}) \equiv \max_{1 \leq j \leq m} S(\boldsymbol{y}, \boldsymbol{d}_j) \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{6.6}$$

$$\pi_0 F_0(t) = \begin{cases} F(t), & \text{pour } t \leq \mu_0, \\ \pi_0 - G(t), & \text{pour } t > \mu_0. \end{cases} \tag{6.7}$$

$$\overline{F}(\mu) = \overline{G}(\mu) \tag{6.8}$$

$$\widehat{\mu}_0 = \frac{t_{(n)} + t_{(n+1)}}{2} \tag{6.9}$$

$$\widehat{\pi}_0 = \min\{2n_0/n, 1\} \tag{6.10}$$



$$\widehat{F}_0(t) = \frac{\#\{s_{0,i} \leq t\} + \#\{g_{0,i} \leq t\}}{2n_0} \tag{6.11}$$

$$p_i = 1 - \widehat{F}_0\left(T_{\max}(\boldsymbol{y}_i)\right), \quad \text{pour } 1 \leq i \leq n \tag{6.12}$$

$$\widehat{t}_q = \min\left\{t \geq 0 : \frac{1 + \#\{w_j \leq -t\}}{1 \vee \#\{w_j \geq t\}} \leq q\right\} \tag{6.13}$$

$$w_i = w_i^+ \vee w_i^- \times \begin{cases} +1 \text{ si } w_i^+ > w_i^-, \\ -1 \text{ si } w_i^+ < w_i^-, \end{cases} \tag{6.14}$$

$$p_i = 1 - \widehat{F}_0\left(w_i^+\right) = \widehat{G}_0(w_i^+), \quad \text{pour } 1 \leq i \leq n. \tag{6.15}$$

$$\widehat{q}_k = \frac{1 + \#\{i \in \mathcal{A}_k, w_i < 0\}}{1 \vee \#\{i \in \mathcal{A}_k, w_i > 0\}}. \tag{6.16}$$

$$w_i = \boldsymbol{d}^T \boldsymbol{y}_i \tag{6.17}$$

$$\alpha_m = \Pr\left(\max \boldsymbol{z}^m > \eta\right), \qquad \text{sous } \mathcal{H}_0. \tag{6.18}$$

$$\begin{aligned} \alpha_m &= 1 - \Pr\left(\max \boldsymbol{z}^m \leq \eta\right) = 1 - \Pr\left(z_1^m \leq \eta\right)^m, \\ &= 1 - \Phi\left(\eta\right)^m, \end{aligned} \tag{6.19}$$

$$M_{m+1}(t) = \Pr\left(z_1^{m+1} \leq t \mid z_2^{m+1} \leq t, z_3^{m+1} \leq t\right) \times M_m(t), \tag{6.20}$$

$$\alpha_m = \Pr\left(\max \boldsymbol{z}^m > \eta\right), \qquad \text{under } \mathcal{H}_0. \tag{6.18}$$

$$\begin{aligned} \alpha_m &= 1 - \Pr\left(\max \boldsymbol{z}^m \leq \eta\right) = 1 - \Pr\left(z_1^m \leq \eta\right)^m, \\ &= 1 - \Phi\left(\eta\right)^m, \end{aligned} \tag{6.19}$$

$$M_{m+1}(t) = \Pr\left(z_1^{m+1} \leq t \mid z_2^{m+1} \leq t, z_3^{m+1} \leq t\right) \times M_m(t), \tag{6.20}$$

$$\Pr\left(z_1^{m+1} \leq t \mid z_2^{m+1} \leq t, ..., z_{m+1}^{m+1} \leq t\right) \geq \Pr\left(z_1^{m+1} \leq t \mid z_2^{m+1} \leq t, z_3^{m+1} \leq t\right) \tag{A.1}$$

$$\Pr\left(z_2^{m+1} \leq t, ..., z_{m+1}^{m+1} \leq t\right) \geq \Pr\left(z_1^m \leq t, ..., z_m^m \leq t\right). \tag{A.2}$$

Chapitre 5

Problématique de détection

Sommaire

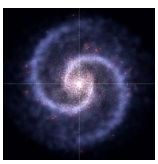
Ce chapitre expose la problématique de détection de sources étendues dans les champs profonds MUSE. Le paragraphe 5.1 décrit le contexte astrophysique de ce problème. Le paragraphe 5.2 présente l'état de l'art des approches de détection en hyperspectral, ainsi que les méthodes développées ces dernières années pour répondre aux problèmes spécifiques des données MUSE. Enfin le paragraphe 5.3 expose les notions de tests d'hypothèses et de tests multiples qui seront exploitées par la suite.

5.1 Contexte astrophysique

L'observation de champs profonds comme l'UDF par MUSE permet de rechercher des structures extragalactiques très peu intenses mais très étendues spatialement, notamment le Circum-Galactic Medium (CGM) puis l'Inter-Galactic Medium (IGM). Le CGM est composé de grandes étendues de gaz (principalement de l'hydrogène) entourant certaines galaxies, sous forme de halos, en interaction forte avec elles.

On peut voir sur la figure 5.1 une simulation des structures extragalactiques reliant les galaxies entre elles. Le CGM et l'IGM sont ainsi les pourvoyeurs d'hydrogène permettant la formation des étoiles au sein des galaxies.

Ces halos de gaz sont si étendus qu'ils peuvent être résolus spatialement par MUSE contrairement aux galaxies en leur cœur. Toutefois leur densité étant très faible, ils sont également très peu lumineux, même dans les longueurs d'onde où ils émettent le plus fortement (raie d'émission Lyman-α), ce qui rend leur détection particulièrement difficile. Cette difficulté est amplifiée par la présence de nombreuses autres sources de nuisance très lumineuses, les galaxies voisines. On peut voir sur la figure 5.2 une galaxie entourée d'un halo, tels qu'observés par MUSE.



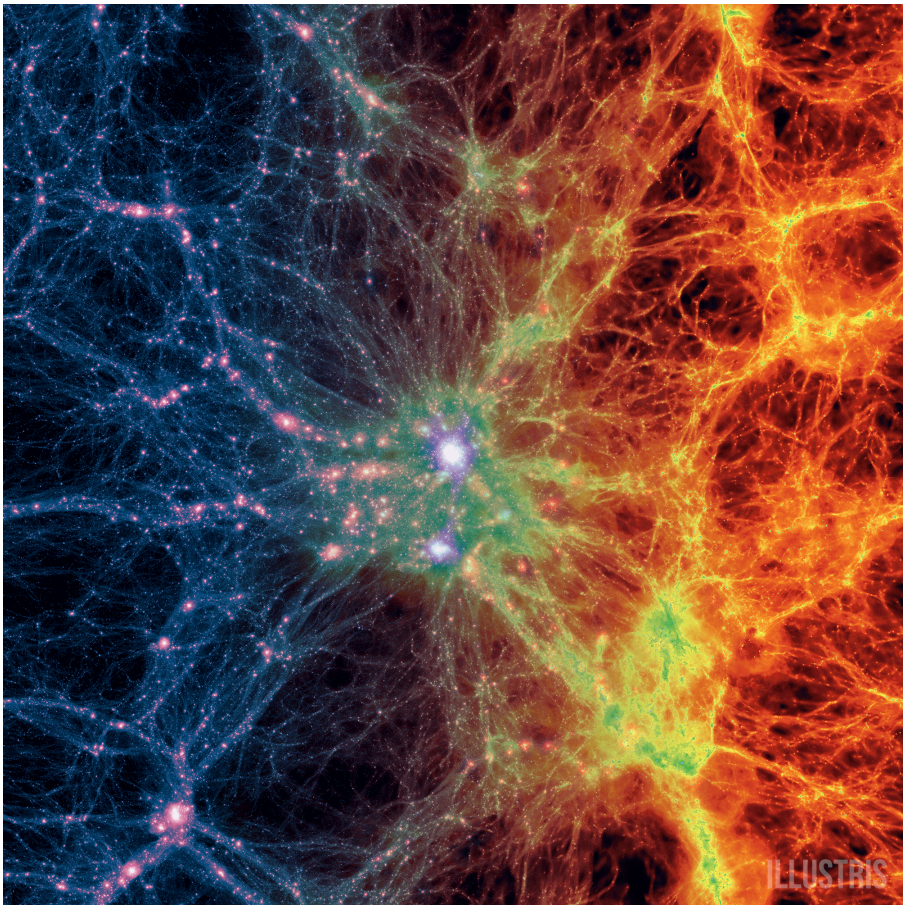
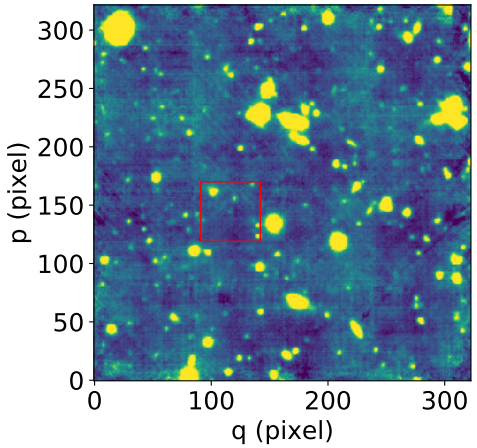
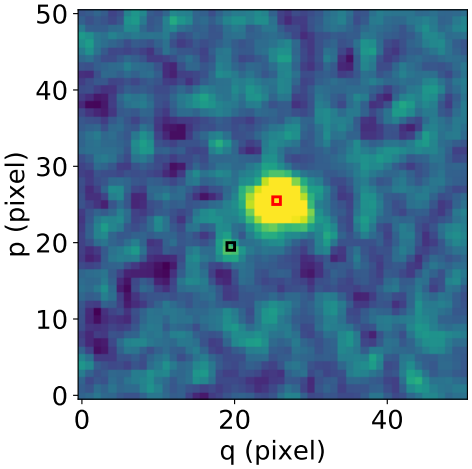


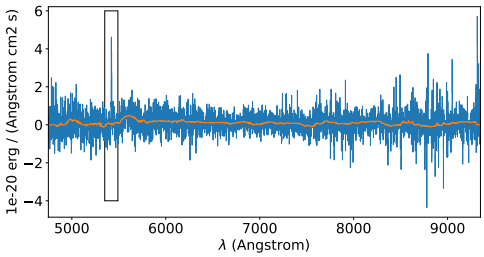
FIGURE 5.1 – Simulation des structures extragalactiques. On peut voir sur la gauche de l'image la distribution de la matière noire, qui va contraindre la structure de la matière classique, visible à droite. Les galaxies apparaissent comme les noeuds de cette toile cosmique. Elles sont en interaction forte avec des halos de gaz proches (le CGM) qui s'éloigne ensuite jusqu'à former les filaments de l'IGM. Crédits : Illustris Simulation.



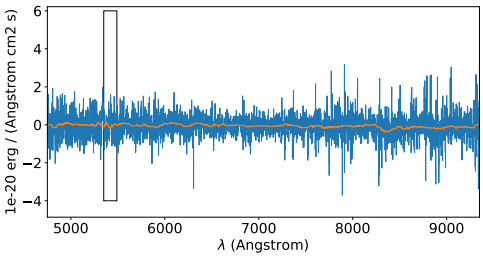
(a) Image blanche MUSE, la région considérée est marquée par le rectangle rouge.



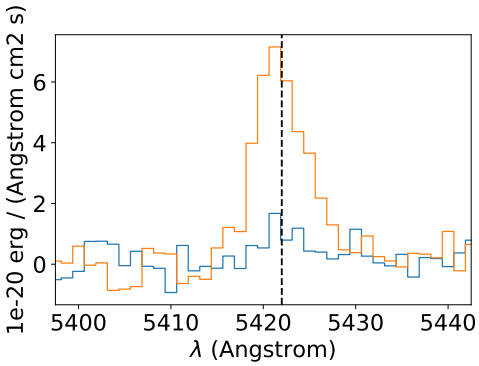
(b) Zoom sur la région considérée, image blanche prétraitée (soustraction continu et filtrage adapté spatial)



(c) Spectre brut pixel central de la galaxie (point rouge sur l'image b). En orange le continuum. La raie Lyman se trouve dans le rectangle noir.



(d) Spectre brut d'un pixel halo (point noir l'image b). En orange le continuum.



(e) Zoom autour de la raie. Spectres prétraités (soustraction continu et filtrage adapté spatial). En orange le spectre du pixel central de la galaxie, en bleu le pixel du halo

FIGURE 5.2 – Comparaison entre le spectre d'un halo et celui de sa galaxie. La galaxie possède la raie Lyman-α et un (faible) continuum spectral, le halo ne possède plus que la raie Lyman-α, qui diminue fortement en intensité plus on s'éloigne de la galaxie.

Rappelons que cette raie Lyman-α peut se trouver n'importe où dans le spectre observé du fait du redshift (voir chapitre 1). Toutefois le halo galactique se trouvant en première approximation à la même distance de nous que la galaxie qu'il entoure, on peut se limiter à explorer le voisinage spectral de la raie Lyman-α détectée sur la galaxie. On s'attend également à ce que les halos circum-galactiques émettent des raies Lyman-α proches spectralement (en première approximation) des raies Lyman-α émises par le coeur des galaxies, ce qui permet d'avoir une connaissance partielle de la signature spectrale du halo recherché. Les halos étant étendus spatialement et de taille a priori inconnue, il va être nécessaire d'explorer une large région environnant une galaxie afin de qualifier la présence puis la forme d'un halo circum-galactique.

Nous considérons donc dans cette étude un problème de détection de cible/source dans un jeu de données massives, impliquant un très grand nombre de tests statistiques. Dans ce contexte, nous verrons dans le paragraphe 5.3 que cette situation de tests multiples nécessite l'utilisation d'une méthode de contrôle global des erreurs, afin de limiter la quantité de fausses détections tout en maintenant une puissance de détection importante.

De plus la présence de nombreuses sources de nuisance, avec de fortes dynamiques d'intensité rend l'estimation de la statistique du fond particulièrement difficile.

En résumé, la détection du CGM se ramène à un problème de détection dont les principales caractéristiques sont :

- La signature des sources est faible, partiellement connue et potentiellement variable.
- Les sources sont spatialement étendues.
- Le fond est difficile à modéliser et de nombreuses sources de nuisances à forte intensité peuvent être présentes.
- La taille des données à tester nécessite la mise en place d'un contrôle robuste et global des erreurs de détection.

Dans le paragraphe suivant, on va donc exposer l'état de l'art des méthodes qui pourraient permettre de répondre à cette problématique.

5.2 État de l'art

5.2.1 Méthodes de détection en hyperspectral

De nombreuses méthodes ont été développées ces dernières années pour la détection de cibles dans des données hyperspectrales (voir par exemple [Manolakis et al. 2009] pour une synthèse), nécessitant pour la grande majorité des informations sur la signature du fond ou des cibles. Parmi ces approches, on peut notamment citer les détecteurs d'anomalies et les détecteurs de motifs spectraux. Les détecteurs d'anomalies permettent de détecter un signal de nature inconnue, en s'appuyant sur une modélisation paramétrique statistique de la signature du fond (Reed et Yu 1990). Les détecteurs de motifs spectraux comme les approches par filtrage adapté (Manolakis et al. 2000) ou la méthode ACE (*adaptive cosine estimators*) [Scharf et McWhorter 1996], reposent à la fois sur une connaissance *a priori* de la signature de la cible et sur une caractérisation du fond. La plupart de ces approches ne fournissent toutefois pas un contrôle global du taux de fausses alarmes. De plus ces détecteurs ont été développés dans le cadre de la télédétection (Solomon et Rock 1985) avec des RSB élevés et ne sont pas aisément adaptables aux défis proposées par les données MUSE (faible RSB, absence de vérité terrain de calibration).

Un autre ensemble de méthodes s'appuie sur la représentation parcimonieuse (Chen et al. 2011,Bourguignon et al. 2011) à l'aide d'un dictionnaire d'apprentissage. Ces approches sont le plus souvent issues de méthodes de reconstruction, exploitées à des fins de détection. A notre connaissance, ces méthodes ne permettent pas de calibrer le contrôle des erreurs de type I (fausses alarmes), sauf à utiliser des bases de données d'apprentissage non disponibles dans notre contexte astrophysique.

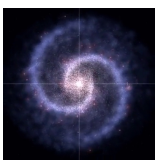
5.2.2 Tests d'hypothèses par maximum de vraisemblance

D'autres approches récentes (Paris et al. 2013 ; Meillier 2015 ; Courbot et al. 2017b) se sont intéressées à la détection de sources faibles dans le contexte des données MUSE à l'aide de stratégies par tests d'hypothèses. En particulier, la méthode proposée dans [Courbot et al. 2017b] s'appuie sur un test de vraisemblance généralisé (GLR) en deux temps. Une première détection du coeur brillant de la galaxie est effectuée, permettant de sélectionner un ensemble d'atomes spectraux pertinents. Une deuxième détection par GLR, avec une contrainte de parcimonie sur les atomes du dictionnaire sélectionné, est ensuite appliquée afin de détecter les extensions spatiales peu lumineuses. Cette approche nécessite toutefois une modélisation adéquate du bruit, notamment de sa matrice de variance-covariance. Un contrôle précis des erreurs de détection est alors difficile à assurer lorsque le bruit et ses structures de dépendances s'éloignent du modèle considéré. Des approches de type max-test ont été également été développées récemment (Meillier 2015) mais dans le but de servir de processus de présélection alimentant une procédure par processus ponctuel marqué.

5.2.3 Détection par classification/segmentation

Une autre approche permettant la détection de sources étendues est d'aborder ce problème comme un problème de classification ou de segmentation. Des travaux exploratoires ont d'ailleurs été réalisés en ce sens durant cette thèse et sont présentés en annexe C Ce type d'approches a toutefois notamment été développé dans [Courbot et al. 2016] à l'aide de champs de Markov triplets orientés, et dans [Courbot et al. 2017a] avec l'utilisation d'arbres de Markov (pour de plus amples détails, le lecteur peut également se référer à [Courbot 2017]). L'intérêt de ce type d'approche est de pouvoir prendre en compte toutes les informations de structures spatiales possibles et s'inscrivent dans le cadre plus large de la détection des filaments extragalactiques de l'IGM, avec des résultats obtenus très prometteurs. Là encore, il est difficile de définir et de contrôler un niveau d'erreur donné, bien que ces méthodes fournissent un niveau d'incertitude sur le résultat de classification de chaque pixel. Notre objectif ici est *a contrario* d'assurer notamment un contrôle fort des erreurs, robuste aux erreurs de modélisation du bruit (en limitant fortement le nombre d'hypothèses faites sur la distribution de ce bruit). Notons par ailleurs, qu'au vu de la faiblesse des signaux recherchés, une pluralité de méthodes fonctionnelles ne peut qu'être bénéfique à l'application astrophysique sous-jacente.

La méthode proposée par la suite va ainsi s'appuyer sur plusieurs concepts (tests d'hypothèses, approches parcimonieuses, FDR) qui sont brièvement rappelés dans les paragraphes suivants.



5.3 Test d'hypothèses

On rappelle ici quelques notions sur les tests d'hypothèses statistiques utilisés par la suite. Un test d'hypothèses statistiques est une méthode d'inférence statistique permettant d'évaluer une (ou plusieurs) hypothèse(s) statistique(s) en fonction d'un jeu de données. Dans le cadre de la détection, on définit le plus souvent deux hypothèses, l'une (*hypothèse nulle*) correspondant à la classe des signaux non désirés (bruit, fond) et l'autre (*hypothèse alternative*) correspondant au signal recherché. Le choix parmi l'hypothèse nulle et l'hypothèse alternative est alors caractérisé par deux type d'erreurs : l'erreur de type I est la probabilité de rejeter à tort un échantillon sous hypothèse nulle, et l'erreur de type II est la probabilité de considérer un échantillon comme nul alors qu'il appartient en vérité à l'hypothèse alternative. Dans l'approche classique de Neyman-Person, on cherche alors à contrôler lors d'un test l'erreur de type I. On peut voir sur la figure 5.3 la définition des erreurs et de la puissance d'un test en fonction des distributions des statistiques de test sous l'hypothèse nulle (première ligne) et l'hypothèse alternative (seconde ligne). Il est ainsi à noter que l'erreur de type I est définie (et donc contrôlable) uniquement à l'aide de la distribution sous l'hypothèse nulle.

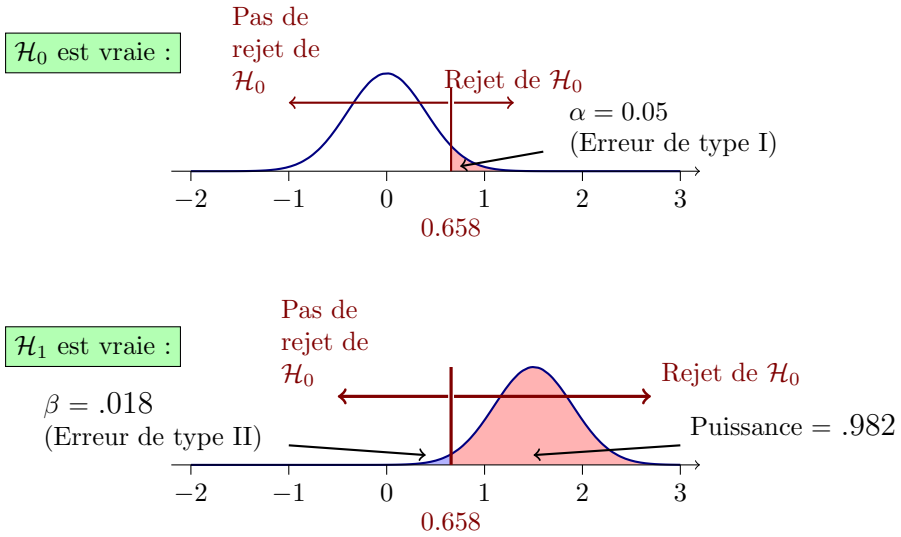


FIGURE 5.3 – Caractérisation d'un test d'hypothèses. On fixe le seuil à 0.658. Sur la première ligne, les échantillons sont distribués selon l'hypothèse nulle, et selon l'hypothèse alternative sur la seconde ligne.

On définit alors la p-valeur comme la probabilité d'obtenir la même valeur (ou une valeur encore plus extrême) du test si l'hypothèse nulle était vraie : par exemple dans le cadre d'un test unilatéral, pour un échantillon i de statistique de test t_i, la p-valeur p_i vaut $p_i = \Pr_{\mathcal{H}_0}(t \geq t_i)$. Plus la p-valeur d'un test est proche de zéro, moins il est probable que l'échantillon testé se trouve sous l'hypothèse $\mathcal{H}_0$.

5.3.1 Tests multiples

Comme énoncé en section 5.1, une contrainte majeure de la détection du CGM est d'assurer un contrôle significatif des erreurs de détection dans le contexte d'un très grand nombre de tests simultanés. On parle alors de tests multiples.

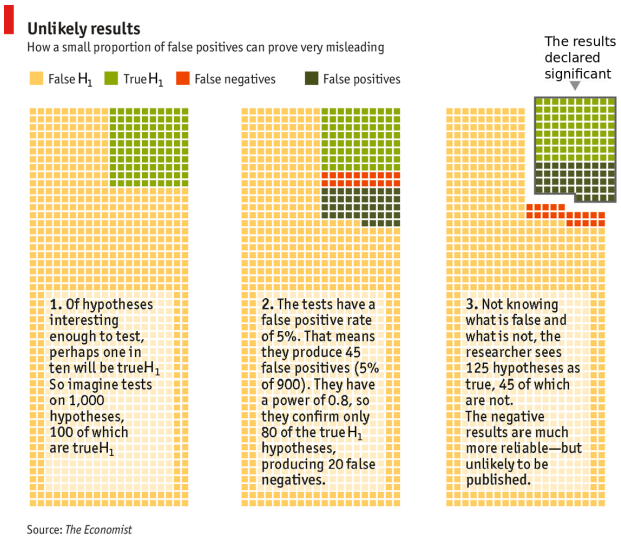


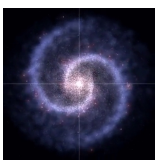
FIGURE 5.4 – Exemple extrait d'un article de *The Economist* en 2013, intitulé "Unreliable research" : si la puissance du test est de seulement 0.4, on obtiendra dans cet exemple 40 vrais positifs en moyenne, pour 45 faux positifs, ce qui interroge sur la significativité des détections.

Dans ce contexte, le contrôle du risque d'erreur de type I s'avère inapproprié (GENOVESE et al. 2002 ; MEILLIER et al. 2015) : le nombre d'hypothèses nulles rejetées à tort (fausses alarmes) peut rapidement devenir important (i.e. jusqu'à dépasser le nombre de bonnes détections) du fait du grand nombre de tests réalisés. En effet, notons T la statistique de test et $\mathcal{R}_\alpha$ la région de rejet au niveau α : si $\mathcal{H}_0$ est vraie, $\Pr(T \in \mathcal{R}_\alpha) = \alpha$. Si on considère alors N statistiques indépendantes $T_1, \dots, T_N$ toutes obtenues sous l'hypothèse nulle $\mathcal{H}_0$, la probabilité de rejeter, à tort, *au moins une* des N hypothèses nulles est de :

$$\begin{aligned}\Pr\left(\exists T_i \in \mathcal{R}_\alpha\right) &= 1 - \Pr\left(T_1, \dots, T_N \notin \mathcal{R}_\alpha\right) = 1 - \prod_{i=1}^{N} \Pr\left(T_i \notin \mathcal{R}_\alpha\right), \\ &= 1 - \prod_{i=1}^{N} (1-\alpha) = 1 - (1-\alpha)^N\end{aligned}$$

Ainsi pour un niveau de contrôle classiquement utilisé de $\alpha = 0.05$, effectuer $N = 20$ tests donne une probabilité de 0.64 de trouver une découverte "significative" par pur hasard. En effet la probabilité d'avoir **au moins** un faux positif est bien supérieure à la probabilité que le $i$ème test soit un faux positif. Ce phénomène est également illustré par un exemple extrait du journal *The Economist* en 2013, présenté dans la figure 5.4, qui montre le manque de pertinence d'un contrôle individuel des erreurs en situation de tests multiples. Citons enfin les travaux de [IOANNIDIS 2005] montrant que de nombreuses découvertes, publiées dans des journaux de prestige, sont difficilement reproductibles, notamment du fait de cette non prise en compte des situations de tests multiples.

On le voit, la multiplication des tests sur des données de grande taille pose des interrogations majeures sur la significativité des "découvertes". Pour pallier ces problèmes, on cherche donc à définir et contrôler un critère global des erreurs de type I.



5.3.2 Besoin d'un contrôle global : le FDR

Un des premiers critères introduits pour prendre en compte cette problématique de tests multiples est le *family-wise error rate* (FWER) qui consiste à contrôler la probabilité de faire *au moins une* erreur de type I en réalisant une série de tests. Ce FWER peut être contrôlé par différentes approches, notamment la procédure Bonferroni (Dunn 1961) : notons $\mathcal{H}_1, \dots, \mathcal{H}_N$ une famille d'hypothèses et $p_1, \dots, p_N$ les p-valeurs correspondantes ; la correction de Bonferroni consiste à rejeter l'hypothèse nulle pour chaque $p_i \leq \frac{\alpha}{N}$, contrôlant ainsi le FWER à un niveau α.

Ce contrôle ne nécessite aucune hypothèse sur la dépendance entre les p-valeurs ou sur le nombre de vraies hypothèses nulles. Le critère FWER est toutefois très conservatif et n'est donc pas adapté à la détection de sources faibles.

Benjamini et Hochberg proposent alors dans [Benjamini et Hochberg 1995] de substituer au contrôle individuel de l'erreur de type I (ou contrôle de la probabilité de fausse alarme) une mesure du contrôle global des erreurs, nommée *False Discovery Rate* (FDR), ainsi qu'une procédure de contrôle de ce FDR. On introduit pour cela une nouvelle terminologie autour de la notion de découverte :

- $R \equiv \#$ Découvertes (Rejets de $\mathcal{H}_0$)
- $U \equiv \#$ Fausses découvertes (Faux positifs),
- $T \equiv \#$ Vraies découvertes (Vrais positifs),

On définit ainsi la proportion de fausses découvertes parmi toutes les découvertes, appelée FDP (*False Discovery Proportion*) :

$$\mathrm{FDP} \equiv \frac{U}{R \vee 1},$$

où $R \vee 1$ indique le maximum entre R et 1 (ce qui permet ici simplement de garantir la validité de la formule en l'absence de découvertes). Cette proportion est une variable aléatoire. Benjamini et Hochberg proposent de la contrôler en moyenne en introduisant le taux de fausses découvertes, le FDR (*False Discovery Rate*) :

$$\mathrm{FDR} \equiv \mathbb{E}\left[\mathrm{FDP}\right] = \mathbb{E}\left[\frac{U}{R \vee 1}\right]. \tag{5.1}$$

La table 5.1 récapitule les différentes quantités définies lors de tests. On peut remarquer que les taux d'erreurs individuelles et la puissance sont calculées horizontalement alors que le FDR est calculé verticalement, de manière plus similaire à l'interprétation bayésienne des tests d'hypothèses (où on cherche à savoir si l'hypothèse $\mathcal{H}_0$ est vraie ou fausse sachant les données observées, voir par exemple [Storey 2003]).

Plusieurs méthodes peuvent être mises en place pour contrôler le FDR. La plus classique est celle accompagnant la définition même du FDR dans [Benjamini et Hochberg 1995]. Cette procédure nécessite de connaître les p-valeurs du test (et donc pour cela la distribution sous l'hypothèse nulle). Cette procédure assure un contrôle exact et est très couramment utilisée, notamment du fait de sa simplicité d'exécution. Toutefois, elle n'est pas nécessairement la plus performante en terme de puissance de détection. De plus, dans le cadre de données réelles, la distribution du bruit est rarement parfaitement connue. Par exemple lorsque le bruit présente une structure de dépendance complexe, éventuellement hétéroscédastique, un estimateur robuste de la distribution à partir des observations peut être difficile voir impossible à obtenir.

		Décision		
		H_0 conservés	H_0 rejetés	Total
Réalité	Vrais $\mathcal{H}_0$	V	U	N_0
	Faux $\mathcal{H}_0$	S	T	N_1
	Total	$N-R$	R	N

TABLE 5.1 – Récapitulatif des différentes quantités définies lors de tests. Le contrôle classique individuel des erreurs de type I se fait de façon horizontal (contrôle en moyenne de $\frac{U}{N_0}$) alors que le contrôle FDR se fait verticalement (contrôle en moyenne de $\frac{U}{R}$).

Des travaux récents (BARBER et CANDÈS 2015 ; BARBER et RAMDAS 2015) ont ainsi porté sur de nouvelles approches de contrôle du FDR, basées sur la constructions de contrefaçons puis de statistiques de contrôle, afin d'améliorer la puissance de détection des tests.

5.3.3 Approches parcimonieuses

La signature des cibles recherchées n'est que partiellement connue : on s'attend à ce que les halos circum-galactiques émettent des raies Lyman-α proches spectralement (en première approximation) des raies Lyman-α émises par le cœur des galaxies. Afin de caractériser cette signature on peut donc chercher à la décomposer de façon parcimonieuse sur un dictionnaire (appris par exemple sur les raies des galaxies).

Les problèmes des détection d'événements rares ou de signaux parcimonieux ont été traités abondamment dans la littérature récente. Les procédures à comparaisons multiples comme le *higher criticism* (HC) et les approches de type Bonferroni, ont notamment été proposées et ont montré avoir des propriétés asymptotiques optimales sous des régimes parcimonieux (DONOHO et JIN 2004 ; HALL et JIN 2010 ; ARIAS-CASTRO et al. 2011 ; DONOHO et JIN 2015). De telles méthodes ne nécessitent pas la spécification des événements/signaux à détecter. Les procédures HC peuvent être vue comme s'adaptant à l'intensité et au degré inconnu de parcimonie des signaux à détecter.

Ces procédures par comparaison multiples peuvent notamment s'appliquer sur les représentations des signaux sur des dictionnaires. En particulier, des dictionnaires saturés et/ou cohérents sont susceptibles de fournir des représentations parcimonieuses adaptées à l'application souhaitée, et il a été montré qu'ils pouvaient grandement améliorer la puissance de détection des tests (voir par ex. DONOHO et JIN 2004). Nous allons donc chercher à représenter de façon parcimonieuse sur un dictionnaire la signature spectrale recherchée et nous appuyer par la suite sur une stratégie de détection d'événements rares, le max-test, pour établir notre test de détection.

Résumé

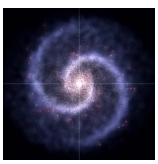
Dans ce chapitre nous avons exposé la problématique de détection du circum-galactic medium, structure de gaz entourant les galaxies, qui peut se caractériser comme la détection de source étendue, de signature partiellement connue (i.e. décomposable de façon parcimonieuse sur un dictionnaire) dans un fond et un bruit difficilement modélisable.
Nous avons ensuite exposé l'état de l'art, qui se décompose principalement en deux catégories : les travaux de détection hyperspectrale classique, peu adaptés aux données MUSE, et les travaux plus récents s'intéressant à la problématique de MUSE, qui manquent généralement d'un contrôle robuste des erreurs. Nous avons finalement démontré la nécessité d'un contrôle global des erreurs en situation de tests multiples.
Dans le chapitre suivant, nous allons désormais développer des méthodes permettant de répondre à ce problème de détection, en s'appuyant sur des tests d'hypothèses multiples, en assurant un contrôle global des erreurs de type FDR.

Chapitre 6

Méthode de détection

Sommaire

Ce chapitre expose la méthode proposée pour détecter le signal d'intérêt tout en garantissant un contrôle global des erreurs. Le paragraphe 6.1 formalise le problème de détection et expose la construction des statistiques de test. Cette construction s'appuie sur une approche de détection de motifs sur un dictionnaire à forte cohérence pour permettre de prendre en compte une éventuelle variabilité de la cible. Cette méthode repose sur l'étude du max-test conduite dans [Arias-Castro et al. 2011] pour la détection d'événements rares, grâce à une représentation parcimonieuse du signal recherché. Afin de permettre la calibration de la procédure de test pour un contrôle du FDR qui soit robuste aux erreurs de modélisation du fond, une première méthode de contrôle est proposée dans le paragraphe 6.2, s'appuyant sur la procédure de Benjamini et Hochberg (Benjamini et Hochberg 1995). Cette méthode nécessite uniquement comme principale hypothèse que le bruit soit distribué selon une loi symétrique. Elle s'appuie sur un apprentissage, plus précisément une estimation empirique, de la loi du bruit. Elle ne permet toutefois pas de prendre en compte les structures spatiales des cibles recherchées. Pour lever cette limitation, une seconde approche, s'appuyant toujours sur l'hypothèse de symétrie du bruit, est développée dans le paragraphe 6.3 afin d'améliorer la puissance de détection de sources spatialement structurées. La construction d'un dictionnaire adapté aux données MUSE est discutée dans le paragraphe 6.4.



Notations

Par la suite un "pixel" réfère selon le contexte à une position dans la grille spatiale de MUSE ou au vecteur spectral associé. Les vecteurs de caractéristiques associés à un pixel (un spectre dans un contexte hyperspectral) sont représentés par des caractères en gras ($\boldsymbol{y}$). La notation $a \vee b$ indique $\max(a, b)$. La notation $\text{sgn}(a)$ indique le signe de a.

6.1 Construction des statistiques de test

6.1.1 Problème de détection

Nous nous intéressons ici à la détection d'un signal $\boldsymbol{x}$, à partir de données bruitées $\boldsymbol{y} \in \mathbb{R}^l$. Soit $\mathcal{H}_0$ et $\mathcal{H}_1$ les hypothèses représentant respectivement l'absence et la présence de la contribution de la source $\boldsymbol{x}$. Le problème de détection s'écrit alors comme un choix entre les deux hypothèses :

$$\begin{cases} \mathcal{H}_0 : \boldsymbol{y} = \boldsymbol{\epsilon}, \\ \mathcal{H}_1 : \boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\epsilon}, \end{cases} \tag{6.1}$$

où $\boldsymbol{\epsilon} \in \mathbb{R}^l$ est un vecteur de bruit centré et indépendant de $\boldsymbol{x}$, de distribution inconnue.

Lorsque $\boldsymbol{x}$ n'est pas parfaitement connu, une approche classique pour aborder le problème (6.1) consiste à modéliser $\boldsymbol{x}$ comme une combinaison parcimonieuse de signaux de référence, pris dans un dictionnaire redondant $\boldsymbol{D}$. On peut ainsi citer les travaux de [Mallat et Zhang 1993], ou [Huang et Aviyente 2007] dans le cadre de tâches de classification. Les signaux de référence ou *atomes*, correspondent aux vecteurs colonnes $\boldsymbol{d}_j \in \mathbb{R}^l$, pour $1 \leq j \leq m$, de $\boldsymbol{D} \in \mathbb{R}^{l \times m}$, où m est le nombre total de signaux de références. Ces atomes sont classiquement ℓ_2-normalisés : $||\boldsymbol{d}_j||_2 = 1$ pour $1 \leq j \leq m$.

De plus, dans le contexte présent, le signal d'intérêt $\boldsymbol{x}$ est supposé non-négatif. Ainsi, pour assurer une décomposition non-négative, les atomes $\boldsymbol{d}_j$, pour $1 \leq j \leq m$, sont également choisis non-négatifs. Il est à noter que, contrairement à la plupart des techniques de représentations parcimonieuses présentes dans la littérature, nous ne cherchons pas à définir un dictionnaire optimal pour la reconstruction et l'estimation, mais un dictionnaire permettant la définition d'un test de détection efficace. La construction de ce dictionnaire, s'appuyant sur des *a priori* physiques liés à notre application, est abordée dans le paragraphe 6.4.

Sous les hypothèses de non-négativité et de parcimonie, le signal cible peut s'exprimer comme :

$$\begin{aligned} &\boldsymbol{x} \approx a_{i_1}\boldsymbol{d}_{i_1} + \ldots + a_{i_k}\boldsymbol{d}_{i_k}, \\ &\text{avec } a_{i_j} > 0, \quad \text{pour } 1 \leq j \leq k, \end{aligned} \tag{6.2}$$

où $k \ll m$. A partir de cette représentation, le problème de détection se réduit alors à la procédure de test unilatéral (6.3) :

$$\begin{cases} \mathcal{H}_0 : a_1 = a_2 = \ldots = a_m = 0, \\ \mathcal{H}_1 : \text{ au moins un } a_i > 0, \end{cases} \tag{6.3}$$

6.1.2 Statistiques de test

On cherche à présent à définir une statistique de test adaptée au problème de détection (6.3) obtenu par représentation parcimonieuse de $\boldsymbol{x}$. Considérons dans un premier temps la statistique

de test pour un seul atome. Notons $S(\boldsymbol{y}, \boldsymbol{d})$ une mesure de similarité entre les données observées $\boldsymbol{y} \in \mathbb{R}^l$ et un vecteur de référence normalisé $\boldsymbol{d} \in \mathbb{R}^l$.

Parmi les mesures de similarité classiques, on peut notamment considérer la statistique du filtre adapté

$$S_{FA}(\boldsymbol{y}, \boldsymbol{d}) \equiv \left\langle \frac{\boldsymbol{d}}{||\boldsymbol{d}||}, \boldsymbol{y} \right\rangle = \boldsymbol{d}^T \boldsymbol{y}, \tag{6.4}$$

ou bien la distance spectrale angulaire (*spectral angular distance* ou SAD)

$$S_{SAD}(\boldsymbol{y}, \boldsymbol{d}) \equiv \frac{\langle \boldsymbol{d}, \boldsymbol{y} \rangle}{||\boldsymbol{d}||.||\boldsymbol{y}||} = \frac{\boldsymbol{d}^T \boldsymbol{y}}{||\boldsymbol{y}||}, \tag{6.5}$$

qui est très fréquemment utilisée en analyse hyperspectrale (Schowengerdt 2006). Pour un signal donné d'amplitude $a = ||\boldsymbol{y}|| > 0$, ces mesures de similarités sont maximales lorsque $\boldsymbol{y} = a\boldsymbol{d}$.

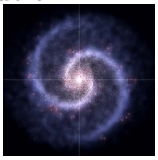
On construit à présent les mesures de similarité $S(\boldsymbol{y}, \boldsymbol{d}_j)$ entre les observations $\boldsymbol{y}$ et les atomes du dictionnaire $\boldsymbol{d}_j$, pour $1 \leq j \leq m$. A partir de ces mesures, une procédure de test pour le problème de tests multiples (vis-à-vis du nombre m d'atomes) exposé en (6.3) peut être obtenue par une approche de type Bonferroni : en prenant en compte (6.2), cela nous amène à considérer le max-test unilatéral suivant

$$T_{\max}(\boldsymbol{y}) \equiv \max_{1 \leq j \leq m} S(\boldsymbol{y}, \boldsymbol{d}_j) \underset{H_0}{\overset{H_1}{\gtrless}} \eta, \tag{6.6}$$

où η est un seuil donné. La motivation derrière l'utilisation de ce max-test est double. Tout d'abord, d'un point de vue théorique, dans le cadre de signaux très parcimonieux, la méthode max-test est asymptotiquement aussi efficace que la méthode asymptotiquement optimale, le *higher criticism* (HC), comme démontré dans [Donoho et Jin 2004 ; Arias-Castro et al. 2011]. De plus, dans une situation nombre fini d'échantillons, comme c'est le cas en pratique avec les données MUSE, il a été observé que le max-test est très efficace (Paris et al. 2013 ; « Nonparametric Bayesian extraction of object configurations in massive data »), et en pratique plus puissant que les méthodes HC (Meillier 2015).

Intéressons nous désormais à l'application de ce max-test (6.6) à un grand nombre n de réalisations $\{\boldsymbol{y}_i\}_{1 \leq i \leq n}$. Nous nous trouvons à nouveau dans un contexte de tests multiples mais cette fois-ci vis-à-vis du nombre d'échantillons n. Nous allons donc chercher à contrôler le nombre de fausses détections effectuées lors du test des n pixels.

Afin de fixer le seuil de décision η tout en contrôlant les erreurs de type I, c'est-à-dire les échantillons sous $\mathcal{H}_0$ qui seront détectés à tort comme étant sous $\mathcal{H}_1$, la distribution sous l'hypothèse nulle $\mathcal{H}_0$ de la statistique du max-test $T_{\max}(\boldsymbol{y})$ doit être connue. Dans le cadre d'applications concrètes comme les données MUSE, du fait des processus physiques et des prétraitements (par exemple des étapes d'interpolation et de soustraction du fond), le bruit est spatialement et spectralement corrélé (et non identiquement distribué le long des longueur d'onde), avec une structure de dépendance complexe et inconnue. La distribution de $T_{\max}(\boldsymbol{\epsilon})$, où $\boldsymbol{\epsilon}$ est le vecteur de bruit introduit en (6.1), ne peut pas être aisément modélisée à l'aide d'une distribution paramétrique classique. Toutefois, dans un contexte de tests à grande échelle, nous allons voir dans le prochain paragraphe qu'il devient possible d'estimer cette distribution.



6.2 Contrôle du FDR à l'aide de la procédure BH

6.2.1 Apprentissage de la distribution de l'hypothèse nulle

Considérons désormais les hypothèses suivantes :

A1. *Le vecteur de bruit $\boldsymbol{\epsilon}$ est centré et distribué selon une loi symétrique : $\boldsymbol{\epsilon}$ et $-\boldsymbol{\epsilon}$ ont la même distribution.*

A2. *La mesure de similarité $S(\boldsymbol{y}, \boldsymbol{d})$ utilisée pour construire la statistique de max-test est une fonction impaire des observations $\boldsymbol{y}$, i.e. $S(\boldsymbol{y}, \boldsymbol{d}) = -S(-\boldsymbol{y}, \boldsymbol{d})$ pour tout échantillon $\boldsymbol{y}$ et pour tout atome $\boldsymbol{d}$.*

L'hypothèse A1 faite sur le bruit est relativement faible et assez raisonnable. Elle est, par exemple, vérifiée pour n'importe quelle distribution elliptique centrée, comme des distributions multivariées gaussiennes ou de Student.

Notons également que l'hypothèse A2 est clairement satisfaite pour les statistiques de filtre adapté ou de SAD décrites respectivement en (6.4) et (6.5). Une conséquence directe de ces hypothèses est la propriété clé suivante :

Proposition 1
Si A1 et A2 sont vérifiées, la statistique du max-test $T_{max}(\boldsymbol{y})$ et l'opposée de la statistique des minima $-T_{min}(\boldsymbol{y})$, où $T_{min}(\boldsymbol{y}) \equiv \min_j S(\boldsymbol{y}, \boldsymbol{d}_j)$, sont identiquement distribuées sous l'hypothèse $\mathcal{H}_0$.

Démonstration. Sous l'hypothèse nulle $\boldsymbol{y} = \boldsymbol{\epsilon}$. D'après A1, $T_{\max}(\boldsymbol{\epsilon})$ et $T_{\max}(-\boldsymbol{\epsilon})$ sont identiquement distribuées. De plus,

$$\begin{aligned} T_{\max}(-\boldsymbol{\epsilon}) &= \max_j S(-\boldsymbol{\epsilon}, \boldsymbol{d}_j) = -\min_j \{-S(-\boldsymbol{\epsilon}, \boldsymbol{d}_j)\}, \\ &= -\min_j S(\boldsymbol{\epsilon}, \boldsymbol{d}_j) = -T_{\min}(\boldsymbol{\epsilon}), \end{aligned}$$

où la première égalité de la seconde ligne est due à A2. On en déduit donc que $T_{\max}(\boldsymbol{\epsilon})$ et -$T_{\min}(\boldsymbol{\epsilon})$ suivent la même distribution. □

Dans un cadre de tests à grande échelle (vis-à-vis du nombre n d'échantillons), les statistiques des maxima et des minima, $T_{\max}(\boldsymbol{y})$ et $T_{\min}(\boldsymbol{y})$, sont calculées pour un grand nombre d'observations $\boldsymbol{y}_i$, pour $1 \leq i \leq n$. Notons $\pi_0 \in]0, 1]$ la proportion réelle d'observations $\boldsymbol{y}_i$ distribuées selon l'hypothèse nulle, et $\pi_1 = 1 - \pi_0$ la proportion d'observations distribuées selon l'hypothèse alternative du problème (6.1). Notons $F(t) = \Pr(T_{\max}(\boldsymbol{y}) \leq t)$ la fonction de répartition de la statistique $T_{\max}(\boldsymbol{y})$ des maxima. Cette fonction de répartition peut s'exprimer comme un mélange de deux classes :

$$F(t) = \pi_0 F_0(t) + \pi_1 F_1(t),$$

où F_0 et F_1 indiquent les fonctions de distributions sous, respectivement, l'hypothèse nulle et l'hypothèse alternative. Sous l'hypothèse de non-négativité introduite dans le paragraphe 6.1.2, $T_{\max}(\boldsymbol{y})$ est stochastiquement plus grand sous $\mathcal{H}_1$ que sous $\mathcal{H}_0$, i.e., $F_0(t) > F_1(t)$ pour tout $t \in \mathbb{R}$. Notons μ_0 la médiane des statistiques de test des maxima sous l'hypothèse nulle [1], i.e.,

1. Par soucis de simplicité, les observations sont supposées suivre des distributions continues. Les statistiques de test T sont donc aussi continues, et leur médiane μ est définie par $\Pr(T \leq \mu) = \Pr(T \geq \mu) = \frac{1}{2}$. L'extension à des statistiques discrètes est laissée au lecteur.

$F_0(\mu_0) = \frac{1}{2}$. Nous introduisons à présent l'hypothèse classique de *zero assumption*, que nous qualifierons par la suite de "bruit seul", formulée par Efron dans un autre contexte (Efron 2012, Chap. 6). Cette hypothèse suppose l'existence d'un domaine où seul le bruit est présent, ce qui permet de construire une procédure pour estimer la distribution nulle (voir *Remarque 2* page 91 pour une discussion de cette hypothèse).

A3 (Hypothèse de bruit seul pour F_1)**.** $F_1(t) = 0$ *pour la région* $t \leq \mu_0$ *où les statistiques des maxima sont le plus vraisemblablement sous* $\mathcal{H}_0$*.*

De cette hypothèse, on peut déduire l'expression suivante :

$$F(t) = \pi_0 F_0(t), \quad \text{pour } t \leq \mu_0.$$

De façon similaire, la fonction de survie $G(t) = \Pr\left(-T_{\min}(\boldsymbol{y}) > t\right)$ de l'opposée de la statistique des minima $-T_{\min}(\boldsymbol{y})$ s'écrit

$$G(t) = \pi_0 G_0(t) + \pi_1 G_1(t),$$

où G_0 et G_1 sont les fonctions de survie de $-T_{\min}(\boldsymbol{y})$ sous, respectivement, l'hypothèse nulle et l'hypothèse alternative. D'après l'hypothèse de non-négativité des sources, il vient que $-T_{\min}(\boldsymbol{y})$ est stochastiquement plus petit sous $\mathcal{H}_1$ que sous $\mathcal{H}_0$, soit $G_0(t) > G_1(t)$. Notons par ailleurs que μ_0 est également la médiane de la distribution sous $\mathcal{H}_0$ de $-T_{\min}(\boldsymbol{y})$ puisque $G_0(\mu_0) = 1 - F_0(\mu_0) = \frac{1}{2}$ selon la proposition 1. Cela nous permet d'introduire l'hypothèse "bruit seul" suivante.

A4 (Hypothèse de bruit seul pour G_1)**.** $G_1(t) = 0$ *pour la région* $t \geq \mu_0$ *où la statistique des opposées des minima sont le plus vraisemblablement sous* $\mathcal{H}_0$*.*

On a donc

$$G(t) = \pi_0 G_0(t), \quad \text{pour } t \geq \mu_0.$$

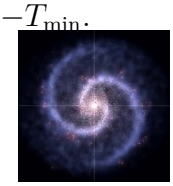
Puisque $G_0(t) = 1 - F_0(t)$ d'après la proposition 1, nous pouvons finalement en déduire l'expression suivante :

$$\pi_0 F_0(t) = \begin{cases} F(t), & \text{pour } t \leq \mu_0, \\ \pi_0 - G(t), & \text{pour } t > \mu_0. \end{cases} \tag{6.7}$$

Le principal intérêt de cette formulation est qu'elle ne dépend plus des distributions de chaque classe mais uniquement de la distribution du mélange. En particulier, les hypothèses A3 et A4 ne nécessitent pas de connaître F_1, ce qui est souvent très compliqué en pratique. L'expression (6.7) est donc robuste à des erreurs de modélisation de l'hypothèse alternative. Cette expression dépend toutefois de la médiane théorique sous l'hypothèse nulle μ_0, de la proportion π_0 d'échantillons sous $\mathcal{H}_0$, et des fonctions de distributions $F(t)$ et $G(t)$, toutes ces quantités étant inconnues. Cependant lorsqu'un grand nombre d'observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ sont disponibles, ces quantités peuvent être estimées à partir des distributions empiriques des statistiques de test. Notons respectivement

$$\overline{F}(t) = \frac{\#\left\{T_{\max}(\boldsymbol{y}_i) \leq t\right\}}{n}, \quad \overline{G}(t) = \frac{\#\left\{-T_{\min}(\boldsymbol{y}_i) > t\right\}}{n},$$

la fonction de distribution empirique de $T_{\max}$ et la fonction de survie empirique de $-T_{\min}$.



A5 (Hypothèse de faible dépendance)**.** *Les fonctions de distribution empiriques* $\overline{F}(t)$ *et* $\overline{G}(t)$ *convergent uniformément vers, respectivement, les fonctions de distribution théoriques* $F(t)$ *et* $G(t)$ *:*

$$\sup_t |\overline{F}(t) - F(t)| \longrightarrow 0, \quad \sup_t |\overline{G}(t) - G(t)| \longrightarrow 0,$$

presque sûrement lorsque le nombre n *d'observations tend vers l'infini.*

La propriété A5 est vérifiée sous condition de faible dépendance entre les échantillons observés $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$. En particulier, elle est valide pour des échantillons indépendants ou avec une dépendance à courte portée d'après le théorème de Glivenko-Cantelli. Une conséquence directe est la convergence simple, en probabilité, de $\overline{F}(t)$ et $\overline{G}(t)$ vers respectivement $F(t)$ et $G(t)$, pour tout $t \in \mathbb{R}$.

Du fait des hypothèses A3 et A4 (hypothèses de bruit seul) et de la proposition 1, μ_0 vérifie $F(\mu_0) = G(\mu_0) = \frac{\pi_0}{2}$. Par conséquent, en s'appuyant sur l'hypothèse A5, un estimateur de la médiane sous l'hypothèse nulle μ_0 peut être cherché comme solution de l'équation en μ suivante :

$$\overline{F}(\mu) = \overline{G}(\mu). \tag{6.8}$$

Lemme 1 (Estimateur empirique de la médiane sous l'hypothèse nulle)
Notons $t_{(1)} < t_{(2)} < \ldots < t_{(2n)}$ *les valeurs ordonnées des statistiques de l'échantillon* $\boldsymbol{t} = (T_{max}(\boldsymbol{y}_1), \ldots, T_{max}(\boldsymbol{y}_n), -T_{min}(\boldsymbol{y}_1), \ldots, -T_{min}(\boldsymbol{y}_n))$*. Notons* $\widehat{\mu}_0$ *la médiane empirique de* $\boldsymbol{t}$*, définie par*

$$\widehat{\mu}_0 = \frac{t_{(n)} + t_{(n+1)}}{2}. \tag{6.9}$$

Alors, lorsque les hypothèses A3 et A4 sont valides, $\widehat{\mu}_0$ *vérifie l'équation* (6.8) *et est un estimateur consistant de la médiane sous l'hypothèse nulle* μ_0*.*

Démonstration. Voir annexe A. □

En s'appuyant sur l'estimateur de la médiane sous l'hypothèse nulle, donné dans le lemme 1, nous pouvons à présent obtenir des estimateurs empiriques de π_0 et $F_0(t)$. Notons

$$\boldsymbol{s}_0 = \{T_{\max}(\boldsymbol{y}_i) \leq \widehat{\mu}_0\},$$

l'ensemble des statistiques du max-test tronquées sur $(-\infty, \widehat{\mu}_0]$, dont les éléments sont notés $s_{0,i}$, pour $1 \leq i \leq n_0$, et où $n_0 = |\boldsymbol{s}_0|$. De façon similaire,

$$\boldsymbol{g}_0 = \{-T_{\min}(\boldsymbol{y}_i) > \widehat{\mu}_0\},$$

indique l'ensemble des statistiques des opposés des minima tronquées sur $(\widehat{\mu}_0, +\infty)$, dont les éléments sont notés $g_{0,i}$ pour $1 \leq i \leq n_0$ (d'après le lemme 1, ces deux ensembles sont de taille identique).

Proposition 2 (Estimateurs empiriques sous $\mathcal{H}_0$)
Sous les hypothèses A3 et A4,

$$\widehat{\pi}_0 = \min\{2n_0/n, 1\}, \tag{6.10}$$

est un estimateur consistent de la proportion π_0 *d'échantillons sous* $\mathcal{H}_0$*, et*

$$\widehat{F}_0(t) = \frac{\#\{s_{0,i} \leq t\} + \#\{g_{0,i} \leq t\}}{2n_0} \tag{6.11}$$

est un estimateur consistent de la distribution sous l'hypothèse nulle $F_0(t)$*, pour* $t \in \mathbb{R}$*.*

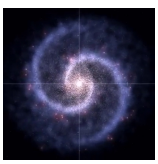
Démonstration. Voir annexe A. □

Remarque 1
La structure de dépendance au sein de chaque observation $\boldsymbol{y}_i \in \mathbb{R}^l$ n'a pas besoin d'être connue pour obtenir les estimateurs empiriques $\widehat{\pi}_0$ et $\widehat{F}_0$ de la proposition 2. Ces estimateurs non-paramétriques reposent principalement sur l'hypothèse A1 de symétrie de la distribution du bruit, qui est peu contraignante. En conséquence, ces estimateurs sont robustes aux erreurs de modèle susceptibles d'apparaître lors de l'utilisation d'estimateurs paramétriques.

Remarque 2
Les hypothèses de "bruit seul" A3 et A4 fournissent un cadre mathématique idéalisé dans lequel les estimateurs empiriques sont consistants. Cependant ces hypothèses sont peu susceptibles d'être parfaitement valides en pratique. En conséquence, l'équation (6.7) est une approximation. Notons toutefois, que plus π_0 est proche de un (autrement dit les cibles recherchés sont peu nombreuses, ce qui est le cas pour un grand nombre d'applications de tests multiples à grande échelle), plus l'approximation sera précise. Cette approximation est également d'autant plus réaliste que $F_0(t)$ (respectivement $G_0(t)$) domine $F_1(t)$ (respectivement $G_1(t)$) pour $t \leq \mu_0$ (respectivement. pour $t \geq \mu_0$). De plus, si quelques observations distribuées selon la loi alternative se trouvent dans les régions où elles sont supposées absentes, alors l'estimateur $\widehat{\pi}_0$ sera biaisé par valeur supérieure, et $\widehat{F}_0(t)$ sera légèrement biaisé vers la distribution alternative $F_1(t)$. Il est donc à noter que du point du vue du contrôle du test statistique à mener, ce léger biais va dans le bon sens : en effet, une procédure de détection basée sur $\widehat{F}_0(t)$ (et $\widehat{\pi}_0$) devient plus conservative car les p-valeurs sont légèrement sur-estimées. ***Cela se traduit bien sûr par une petite perte de puissance mais le contrôle des erreurs de type I est toujours garanti asymptotiquement.***

La figure 6.1 montre les fonctions de densité empiriques associées aux statistiques du max-test $T_{\max}(\boldsymbol{y}_i)$ et à l'opposée $-T_{\min}(\boldsymbol{y}_i)$ des statistiques des minima pour des données synthétiques avec un cadre de test imitant celui de MUSE. Nous pouvons voir que la densité des maxima possède à droite une queue plus lourde que la densité de l'opposée des minima. Cela est dû à la contribution des échantillons sous $\mathcal{H}_1$, la queue à droite de la densité de l'opposée des minima étant elle (approximativement) distribuée selon la densité théorique de l'hypothèse nulle, du fait de l'approximation A4. Par symétrie, la densité des opposés des minima a une queue à gauche plus lourde que la densité des max, puisque c'est là que se manifeste la contribution des échantillons sous $\mathcal{H}_1$.

Afin d'apprécier la précision des estimateurs empiriques de la proposition 2, la densité empirique de l'hypothèse nulle, obtenue à partir de l'échantillon tronqué à droite $\boldsymbol{s}_0$ et de l'échantillon tronqué à gauche $\boldsymbol{g}_0$, est décrite dans la figure 6.2a. Ici la proportion d'échantillons nuls estimée est plus grande que la valeur théorique : $\widehat{\pi}_0 = 0.89$ contre $\pi_0 = 0.81$. Néanmoins, la fonction de densité empirique de l'hypothèse nulle est très proche de la loi théorique. Cela est confirmé par le graphe quantile-quantile entre $\widehat{F}_0$ et la distribution théorique F_0 présenté dans la figure 6.2b. En particulier, cela reste valable dans les queues des distributions, là où la précision de l'estimation des quantiles est cruciale pour la robustesse du test à de faibles niveaux de contrôle.



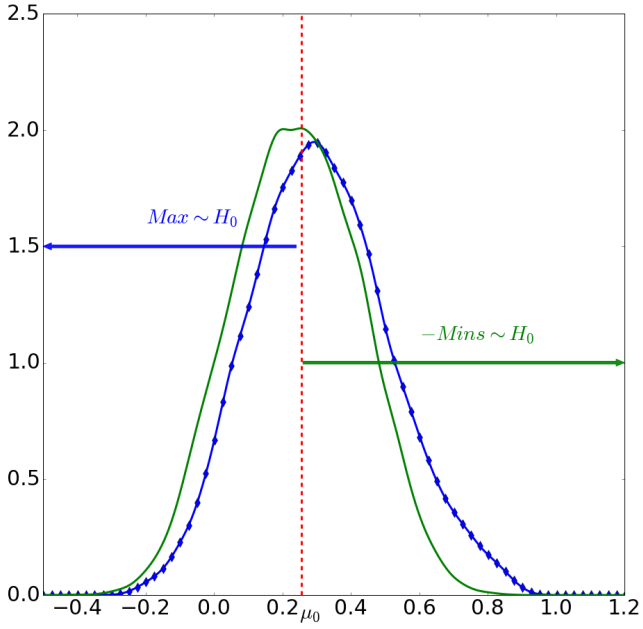
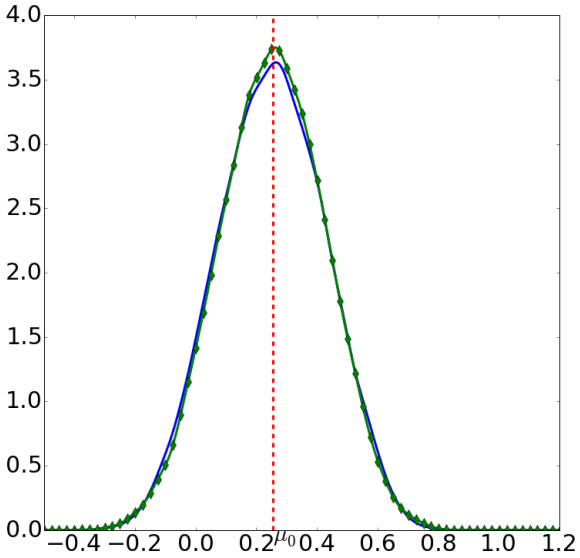
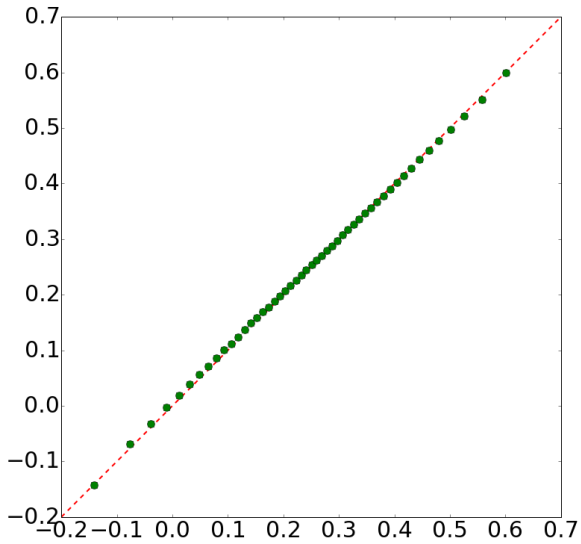


FIGURE 6.1 – Fonctions de densité empiriques des statistiques de max-test $T_{\max}(\boldsymbol{y}_i)$ (courbe bleue avec des marqueurs ♦) et des statistiques des opposées des mins $-T_{\min}(\boldsymbol{y}_i)$ (courbe verte) pour $n = 2500$ échantillons indépendants $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \mathbb{R}^l$ générés à partir du modèle d'observation (6.1) avec $l = 30$. Les entrées du vecteur de bruit $\boldsymbol{\epsilon}_i$ sont i.i.d. et suivent une distribution de Student avec $\nu = 5$ degrés de libertés. La proportion d'hypothèses nulles est $\pi_0 = 0.81$. Les échantillons sous $\mathcal{H}_1$ sont générés par $\boldsymbol{y}_i = a_i \boldsymbol{d} + \boldsymbol{\epsilon}_i$, où $a_i \in [0.1, 3]$ et $\boldsymbol{d}$ est de norme unitaire. Les statistiques des min-tests et max-tests sont obtenues à l'aide d'un dictionnaire $\boldsymbol{D}$ avec $m = 15$ atomes positivement corrélés, dont la vraie signature $\boldsymbol{d}$, et en utilisant la mesure de similarité SAD.

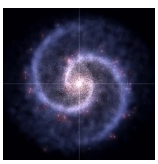


(a) Comparaison de la fonction de densité empirique de l'ensemble $\boldsymbol{t}_0 = \boldsymbol{s}_0 \cup \boldsymbol{g}_0$, où $\boldsymbol{s}_0$ et $\boldsymbol{g}_0$ sont les échantillons utilisés pour construire $\widehat{F}_0(t)$ (courbe bleu) et de la fonction de densité théorique de $T_{\max}(\boldsymbol{y})$ sous $\mathcal{H}_0$ (courbe verte avec des marqueurs ♦).



(b) Courbe quantile-quantile de $\widehat{F}_0(t)$ en fonction de la distribution théorique F_0, indiqué par des marquers • ; la ligne $y = x$ est indiquée en pointillée rouge.

FIGURE 6.2 – Comparaisons entre l'estimation empirique $\widehat{F}_0(t)$ de la distribution sous l'hypothèse nulle et la distribution théorique de $T_{\max}(\boldsymbol{y})$ sous $\mathcal{H}_0$. Les données sont générées avec le même processus que pour la figure 6.1. La fonction de distribution théorique est calculée à l'aide de 10^5 tirages Monte-Carlo.



6.2.2 Contrôle des erreurs

Comme expliqué dans le paragraphe 5.3.1 dans le cadre de tests multiples (plusieurs milliers de pixels pour l'application MUSE), le contrôle classique de l'erreur de type I pour chaque test individuel n'est le plus souvent plus approprié. En effet, le nombre d'hypothèses nulles rejetées à tort peut devenir relativement grand (jusqu'à dépasser parfois le nombre de vraies détections) du fait du grand nombre de tests. Pour faire face à cette difficulté, une approche de contrôle global des erreurs, appelée *False Discovery Rate* ou FDR, a été introduite dans [Benjamini et Hochberg 1995]. Le FDR contrôle la proportion moyenne d'hypothèses nulles rejetées à tort, qualifiés de *fausses découvertes*, parmi toutes les hypothèses nulles rejetées (les *découvertes*) :

$$\mathrm{FDR} = \mathbb{E}\left[\frac{U}{R \vee 1}\right],$$

où R est le nombre total de découvertes (rejet de l'hypothèse nulle), tandis que U est le nombre de fausses découvertes parmi les R découvertes (voir page 82). Une approche simple et largement répandue pour contrôler le FDR est la procédure de Benjamini et Hochberg (ci-après *procédure BH*), également introduite dans [Benjamini et Hochberg 1995]. Notons p_i la p-valeur associée à la $i$ème statistique de test. Notons $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n)}$ les p-valeurs ordonnées, et $\mathcal{H}_0^{(1)}, \ldots, \mathcal{H}_0^{(n)}$ les hypothèses nulles associées à cet ordonnancement. Pour un niveau de contrôle choisi $0 \leq q \leq 1$, la procédure BH_q rejette les hypothèses $\mathcal{H}_0^{(1)}, \ldots, \mathcal{H}_0^{(\hat{k})}$ où

$$\hat{k} = \max\left\{0 \leq k \leq n : \; p_{(k)} \leq q\frac{k}{n}\right\},$$

avec $p_{(0)} = 0$ par convention. Dans notre cadre de détection unilatéral à droite, la $i$ème p-valeur *empirique* peut être obtenue à partir de la distribution empirique sous $\mathcal{H}_0$ exprimée dans la proposition 2 comme

$$p_i = 1 - \widehat{F}_0\left(T_{\max}(\boldsymbol{y}_i)\right), \quad \text{pour } 1 \leq i \leq n. \tag{6.12}$$

Alors, dans le cas de n tests indépendants, ou sous certaines conditions de dépendances positives (Benjamini et Yekutieli 2001), la procédure BH_q contrôle le FDR au niveau $\pi_0 q \leq q$. De plus, si π_0 est connu, la procédure BH peut être appliquée au niveau $\frac{q}{\pi_0}$ pour améliorer la puissance de détection tout en contrôlant le FDR au niveau q. S'appuyant sur cette idée, Storey propose dans [Storey et al. 2004] l'estimateur *modifié* de la proportion nulle π_0 suivant :

$$\hat{\pi}_0^*(\zeta) = \min\left\{\frac{1 + \#\{p_i > \zeta\}}{(1-\zeta)n}, 1\right\}, \quad \text{pour } \zeta \in [0, 1[,$$

où ζ est fixé arbitrairement (classiquement $\frac{1}{2}$). Il est montré dans [Storey et al. 2004] que sous l'hypothèse de faible dépendance A5, la procédure $\mathrm{BH}_{q'}$ appliquée au niveau nominal $q' = q/\hat{\pi}_0^*(\zeta)$ contrôle asymptotiquement le FDR au niveau q. Nous allons voir dans la suite que cette même stratégie peut être utilisée avec l'estimateur empirique $\widehat{\pi}_0$.

Proposition 3 (Estimateur de Storey de π_0)
L'estimateur empirique $\widehat{\pi}_0$ défini dans l'équation (6.10) et l'estimateur de Storey $\hat{\pi}_0^*(\zeta)$ dérivé des p-valeurs empiriques définies dans (6.12) sont égaux pour tout $\zeta = \frac{k}{2n_0}$ avec $k \in \{n_0, \ldots, 2n_0 - 1\}$, et sont asymptotiquement équivalent pour tout $\zeta \in [\frac{1}{2}, 1[$.

Démonstration. Voir annexe A. □

Cette équivalence n'est pas surprenante puisque, comme les estimateurs empiriques proposés ici, l'estimateur de Storey s'appuie sur une hypothèse de bruit seul (autrement dit la fonction de densité des p-valeurs sous $\mathcal{H}_1$ vaut zéro sur $]\zeta, 1]$).

Cela nous amène à considérer la procédure de tests multiples suivante, décrite dans l'algorithme 5.

Algorithme 5 Procédure BH empirique

1: *Entrées :* niveau nominal q du FDR

2: Calcul des estimateurs empiriques sous l'hypothèse nulle $\widehat{\pi}_0$ et $\widehat{F}_0$ définis dans la proposition 2

3: Calcul des p-valeurs empiriques avec l'équation (6.12)

4: Application de la procédure BH avec un niveau nominal de contrôle $q/\widehat{\pi}_0$

5: *Sorties :* liste des pixels détectés

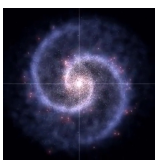
Notons qu'il a été montré dans [Meillier et al. 2015] que dans le cadre de statistiques issues d'un filtrage adapté, avec un signal cible non négatif et sous des hypothèses de bruit i.i.d. gaussien[2], les statistiques de test suivent une condition de *positive regression dependence on a subset* ou PRDS. La procédure BH assure alors un contrôle exact du FDR même avec un nombre fini d'échantillons. Ici le problème est plus complexe. Les statistiques de test sont obtenues à partir de valeurs extrêmes (maxima) qui peuvent être corrélées entre elles. Il est alors difficile d'assurer théoriquement la propriété PRDS, même en ajoutant des hypothèses de gaussianité sur le bruit. Toutefois, sous l'hypothèse de faible dépendance A5, il a été montré dans [Storey et al. 2004] qu'une procédure Oracle, similaire à l'algorithme 5 mais utilisant des p-valeurs obtenues à partir de la distribution F_0 théorique, contrôle asymptotiquement le FDR au niveau q. Comme discuté dans le paragraphe 6.2.1, les p-valeurs empiriques ont tendance à être légèrement biaisées de façon conservative. De plus, si la distribution sous l'hypothèse nulle peut être estimée sur un échantillon plus large que l'échantillon de test, la variance de ces estimées peut être réduite[3]. Cela favorise le contrôle asymptotique pour la méthode proposée.

6.2.3 Validation

La figure 6.3 montre le FDR obtenu avec l'algorithme 5, sur des données simulées 3D (deux dimensions spatiales et une dimension spectrale) soumises à une faible dépendance spatiale (convolution spatiale par un noyau), pour différents niveaux de contrôle nominaux q et pour différents RSB. Le RSB est défini ici par $10 \log \frac{A}{nl\sigma^2}$, où n est le nombre de pixels (c'est à dire, le nombre de tests à effectuer), l est le nombre de bandes spectrales (c'est à dire la dimension des observations $\boldsymbol{y}_i$), σ^2 est la variance marginale du bruit, et $A = ||\boldsymbol{x}||^2$ est l'énergie de la contribution 3D du signal à détecter. Les sources réelles MUSE faibles possèdent des RSB

2. Rappelons que dans le cadre de MUSE, le bruit n'est pas identiquement distribué (car sa variance varie avec la longueur d'onde. L'utilisation directe des résultats de Meillier et al. 2015 nécessiterait donc également de réduire au préalable les données.

3. Cela est notamment possible dans le cadre de notre application : la détection du CGM se fait généralement dans l'environnement proche d'une galaxie, on se limite le plus souvent à explorer une partie seulement d'un jeu de données, mais l'apprentissage peut se faire sur une plus grande portion des données



typiquement de l'ordre de -10dB ou -15 dB. Le protocole expérimental est similaire à celui de la figure 6.1, avec l'ajout de l'application d'une convolution spatiale par un noyau de taille 3 par 3 pixels pour créer des corrélations spatiales. Les estimateurs empiriques définis dans le paragraphe 6.2.1 sont calculés à partir d'une zone étendue spatialement à 200 par 200 pixels (et toujours 30 bandes spectrales).

Cette figure souligne que le contrôle du FDR est bien atteint pour les différents niveaux de RSB. Comme attendu, du fait des approximations A3 et A4, l'algorithme 5 est légèrement plus conservatif que l'Oracle (qui s'appuie sur les vraies F_0 et π_0) à faible RSB, là où la distribution de l'hypothèse alternative est très proche de celle de l'hypothèse nulle (et où donc les approximations A3 et A4 sont moins précises).

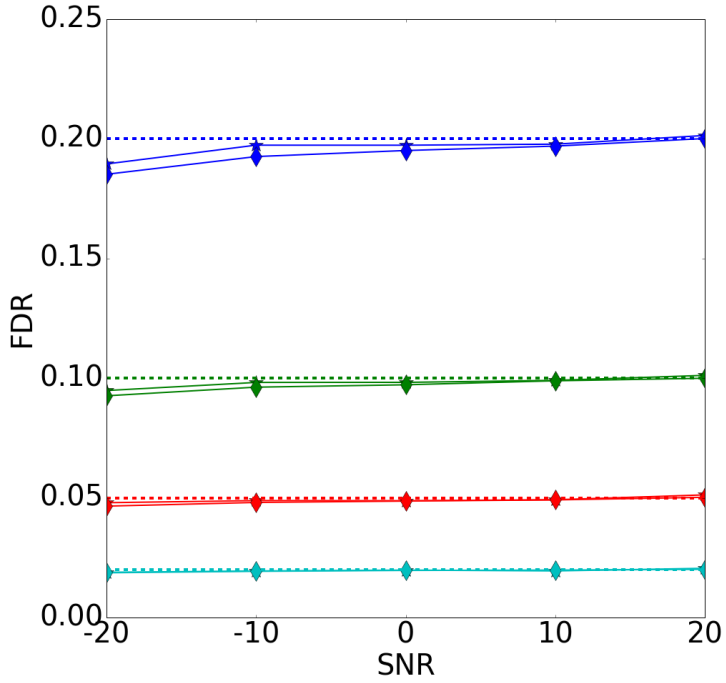


FIGURE 6.3 – FDR empiriques (moyennés sur 1000 tirages Monte-Carlo) en fonction du RSB sous faible dépendance, pour différents niveaux de contrôle du FDR $q = 0.02, 0.05, 0.1, 0.2$ (en cyan, rouge, vert, et bleu, respectivement). Les niveaux nominaux de FDR sont indiqués en pointillés horizontaux ; les FDR empiriques pour l'algorithme FDR sont marqués par des ⧫ ; les FDR empiriques pour la procédure Oracle, basée sur les véritables (i.e. estimées à l'aide de 10^5 tirages Monte-Carlo) valeurs de F_0 et π_0, sont indiqués par des $\star$. Les données sont simulées sous la forme de cubes 3d de dimensions $n = 51 \times 51$ pixels sur $l = 30$ bandes spectrales.

La table 6.1 permet de montrer les avantages d'assurer un contrôle global avec un seuil de détection s'adaptant aux données. Elle compare en effet ce contrôle global avec un contrôle individuel des erreurs basé sur la probabilité de fausses alarmes (PFA). Contrôler à un niveau η (e.g. 5%) la PFA se traduit par la détection de tous les pixels avec une p-valeur inférieure à η. Un tel contrôle par PFA assure alors qu'en moyenne une fraction η de tous les pixels sous $\mathcal{H}_0$ testés seront détectés à tort (mais il ne dit rien de la proportion de ces fausses détections au sein de l'ensemble des pixels détectés).

	PFA 0.05	PFA 0.001	FDR 0.2
Bruit seul			
Fausses détections (pixels)	144	2	0
Vraies détections (pixels)	0	0	0
Source + bruit			
Fausses détections (pixels)	117	2	30
Vraies détections (pixels)	153	106	133
Proportion de fausses découvertes (%)	43.3	1.8	18.4

TABLE 6.1 – Comparaison entre un contrôle du FDR et un contrôle de la PFA sur des données avec et sans cible. Les données sont construites à l'aide de régions de bruit seul dans les données réelles MUSE, dans lesquelles une source synthétique a été ajoutée. Le nombre de pixels testés est de 2500 (images 50 par 50 pixels) et la taille de la source est de 185 pixels. Les résultats sont moyennés sur cinq régions de bruit MUSE différentes.

En présence de données ne contenant que du bruit, la procédure par contrôle de PFA à un niveau nominal de 5% détecte ainsi en moyenne 144 pixels à tort, soit autant que la taille classique d'une source MUSE éventuelle. Afin de s'assurer qu'aucune source n'est détectée à tort, on peut s'orienter vers un niveau plus conservatif par exemple 0.1% ; cela se traduit bien sûr par une puissance de détection plus faible en présence de sources (autour de 55%) alors que le contrôle à 5% permet une bonne puissance (82%), au prix toutefois d'un taux élevé de fausses alarmes (la proportion de fausses découvertes est d'environ 43%). Sur les mêmes jeux de données, un contrôle du FDR à 20% n'entraîne aucune fausse détection en l'absence de source tout en permettant une puissance de détection élevée (72%) en présence de source, montrant ainsi son adaptation à la nature du jeu de données. La table 6.1 montre également que la proportion de fausses découvertes effective est autour de 18% pour un niveau nominal de contrôle du FDR de 20%. Notons que notre procédure nous permet d'estimer le niveau du FDR correspondant à un seuil de détection donné sur les p-valeurs. Par exemple, le seuil appliqué sur les p-valeurs correspondant à une PFA de 5% est équivalent à un niveau FDR théorique d'environ 44% sur les données testées ici. Comme indiqué sur la table 6.1 la proportion de fausses découvertes obtenues pour un contrôle de PFA à 5% est bien autour de 44%.

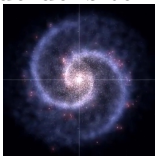
Dans le prochain paragraphe, nous comparons les performances de contrôle du taux d'erreur de la méthode proposée avec une approche de ratio de vraisemblance généralisée (GLR) (inspirée des travaux de [PARIS et al. 2013] et [COURBOT et al. 2017b]). Le bruit est supposé gaussien centré : $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ où la matrice de covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{l \times l}$ est supposée diagonale. Nous avons alors le test de détection suivant :

$$\begin{cases} \mathcal{H}_0 : \boldsymbol{y} = \boldsymbol{\epsilon}, \\ \mathcal{H}_1 : \boldsymbol{y} = \boldsymbol{D}\boldsymbol{a} + \boldsymbol{\epsilon}, \text{ où } ||\boldsymbol{a}||_0 = 1, \boldsymbol{a} \geq \mathbf{0} \end{cases}$$

où $||.||_0$ est la pseudo-norme ℓ_0 (nombre de composantes non-nulles) et $\boldsymbol{a} \geq \mathbf{0}$ est la contrainte de non-négativité sur les coefficients. Le test GLR avec contrainte de 1-parcimonie génère la statistique de test suivante (PARIS et al. 2013) :

$$T_{GLR}(\boldsymbol{y}) = \frac{\max_{\boldsymbol{a}} p(\boldsymbol{y}|\boldsymbol{D}\boldsymbol{a}, \mathcal{H}_1)}{p(\boldsymbol{y}|\mathcal{H}_0)} \text{ avec } ||\boldsymbol{a}||_0 = 1, \boldsymbol{a} \geq \mathbf{0},$$

où $p(\boldsymbol{y}|\boldsymbol{D}\boldsymbol{a}, \mathcal{H}_1)$ indique la fonction de densité de $\boldsymbol{y}$ sous $\mathcal{H}_1$ et $p(\boldsymbol{y}|\mathcal{H}_0)$ la fonction de densité



de $\boldsymbol{y}$ sous $\mathcal{H}_0$. Du fait de l'hypothèse de gaussianité, on en déduit que :

$$T_{GLR}(\boldsymbol{y}) = \frac{\boldsymbol{d}_{\hat{j}}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{y}}{\sqrt{\boldsymbol{d}_{\hat{j}}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{d}_{\hat{j}}}}$$

où $\hat{j}$ est l'indice de la composante non-nulle du vecteur $\hat{\boldsymbol{a}}$ optimal pour la statistique GLR, et $\hat{\boldsymbol{\Sigma}}$ est estimée sur les résidus. Il n'y a pas de forme analytique simple de la distribution de cette statistique puisqu'elle consiste à prendre le maximum d'un vecteur gaussien corrélé. Nous avons donc calibré cette statistique sous $\mathcal{H}_0$ (bruit normal centré) par Monte-Carlo.

Les figures 6.4 et 6.5 illustrent le principal avantage de la méthode proposée : le contrôle est assuré dès lors que la distribution du bruit est symétrique, sans autre spécification. L'approche GLR nécessite une calibration de la distribution sous $\mathcal{H}_0$ donc toute déviation de la loi théorique de $\mathcal{H}_0$ résulte en une perte de contrôle, comme illustré sur la figure 6.5 où le bruit est tiré selon une loi de Student. On peut voir que la procedure BH basée sur les statistiques issues du GLR ne contrôle pas correctement le FDR : dans un premier temps, le FDR réel dépasse fortement le niveau de contrôle donné (permettant au GLR d'être “plus puissant" a un niveau nominal de contrôle donné) ; dans un second temps la méthode devient trop conservative. Ce comportement peut s'expliquer par l'ajustement gaussien de la distribution de Student qui a lieu : les queues sont sous-estimées (d'où l'excès en FDR) alors que le mode est sur-estimé (d'où la perte de puissance dans la seconde partie de la courbe). De plus sur la figure 6.4 on peut voir que lorsque le GLR est le plus à son avantage (modèle choisi correspondant à la vraie distribution du bruit), la méthode proposée reste très proche en terme de puissance tout en étant bien plus polyvalente.

Il est à noter qu'une courbe ROC classique comparant les deux méthodes montrerait des performances très similaires (même puissance pour un même taux effectif d'erreurs) car elle cache l'imprécision de contrôle effectif de l'approche GLR lorsque le modèle n'est pas correctement spécifié. La méthode proposée ne permet pas une plus grande détection à un niveau réel de fausses détections donné, mais assure l'adéquation entre le niveau de contrôle nominal choisi et le contrôle effectif sur les données. Notons enfin que dans le cadre de l'application MUSE, si l'hypothèse de gaussianité est assez bien respectée (voir chapitre 1), la structure de corrélation du bruit doit être estimée avec précision pour assurer un contrôle efficace à l'aide d'une approche GLR. Or, l'estimation de ces covariances est particulièrement difficile, d'autant qu'elles ne sont a priori pas stationnaires (car résultant notamment d'opérations d'interpolation).

6.3 Prise en compte des structures spatiales des sources

Dans le paragraphe précédent nous avons exploité la procédure BH pour effectuer le contrôle du FDR, grâce à un apprentissage robuste de la distribution de la statistique de test des observations de bruit seul. Toutefois, on s'intéresse dans ces travaux à la détection de cibles étendues spatialement, en grande partie connexes. Appliquer la procédure BH dans une approche pixelique ne prend cependant pas en compte les éventuels liens entre les pixels de la cible, ceux-ci pouvant être organisés en structures connexes. La prise en compte d'un tel *a priori* ne peut donc qu'améliorer la puissance de détection. Plusieurs approches par groupe d'échantillons ont été récemment développées par exemple dans les travaux de [Barber et Ramdas 2015] mais s'appuient sur une connaissance *a priori* de ces groupes. Dans notre contexte, la forme de la

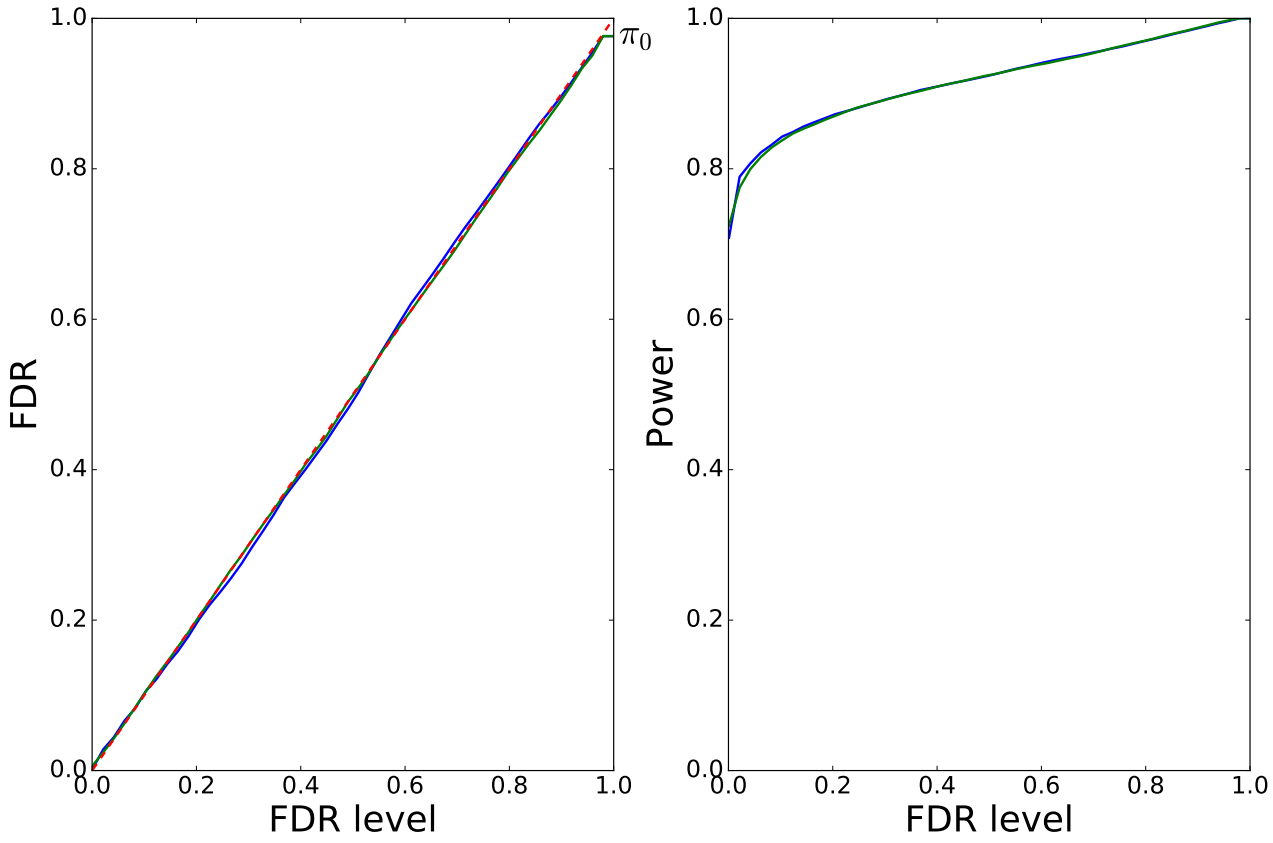
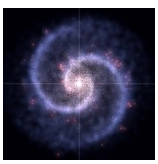


FIGURE 6.4 – Comparaisons des FDR empiriques (gauche) et de la puissance de détection (droite) en fonction du niveau nominal de contrôle FDR, pour le GLR et la méthode proposée, sur des données synthétiques avec un bruit gaussien. Le GLR a été calibré à l'aide de 10^4 tirages de Monte-Carlo sur du bruit normal. La matrice $\mathbf{\Sigma}$ a été estimée sur les données. Les résultats sont moyennés sur 200 tirages de données simulées avec $\pi_0 = 0.97$. Le GLR est en bleu, la méthode proposée est en vert, la droite $y = x$ en pointillé rouge. La mesure de similarité du filtre adapté (6.4) est utilisée pour la méthode proposée.



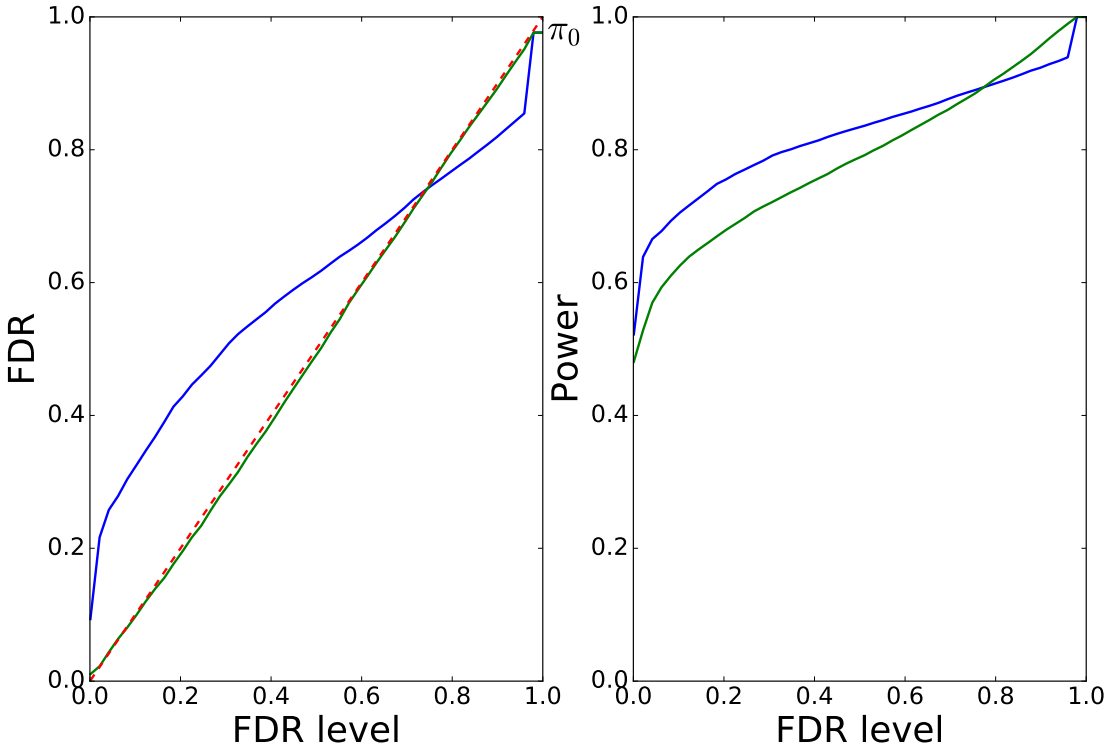


FIGURE 6.5 – Comparaison du FDR effectif (gauche) et de la puissance (droite) en fonction du niveau de contrôle nominal de FDR, entre le GLR et la méthode proposée, sur des données synthétiques avec du bruit de Student (4 degrés de liberté). Le GLR a été calibré à l'aide de 10^4 tirages de Monte-Carlo sur du bruit normal. La matrice $\mathbf{\Sigma}$ a été estimée sur les données. Les résultats sont moyennés sur 200 tirages de données simulées avec $\pi_0 = 0.97$. Le GLR est en bleu, la méthode proposée est en vert, la droite $y = x$ en pointillé rouge. La mesure de similarité du filtre adapté (6.4) est utilisée pour la méthode proposée.

cible est inconnue et nous ne disposons donc pas d'un tel *a priori*. Par ailleurs, un inconvénient notoire des approches classiques de contrôle du FDR est que, par essence, la puissance de détection a tendance à diminuer lorsque le nombre de tests s'accroît : la puissance de détection d'une unique cible, à un niveau de contrôle FDR donné, va donc dépendre de la taille de la région dans laquelle la source est recherchée.

Ainsi, afin de prendre en compte la structure de la cible et de limiter l'influence du nombre de tests, nous développons dans cette section une procédure originale nommée *COnnection accounting Method for Extracting Target* (COMET). COMET s'appuie notamment sur une nouvelle classe d'approches de contrôle du FDR récemment introduite dans [Barber et Candès 2015 ; Candès et al. 2016], que nous proposons d'étendre de sorte à prendre en compte la connexité des structures recherchées.

Reprenons ici le test décrit par (6.1) qui doit être conduit pour un très grand nombre n d'observations $\{\boldsymbol{y}_i\}_{1\leq i\leq n}$. Dans ce contexte de tests multiples, on cherche à déterminer un seuil de décision adapté aux données qui permette de contrôler le FDR. La méthode de contrôle du FDR récemment proposée par Barber et Candès (BC), dans l'étude [Barber et Candès 2015], s'appuie sur la construction de "contrefaçons" (*knockoffs*). Ces contrefaçons sont des variables artificielles reproduisant la structure de corrélation des variables originales, dans un cadre de régression linéaire avec bruit blanc Gaussien. Elles doivent permettre de construire des statistiques de contrôle $\{w_i\}_{1\leq i\leq n}$ pour chaque test, satisfaisant en particulier les propriétés suivantes [4]

- symétrie sous $\mathcal{H}_0$, i.e.
  $\mathbb{P}(w_i > t | i \in \mathcal{H}_0) = \mathbb{P}(w_i < -t | i \in \mathcal{H}_0)$ pour tout $t \in \mathbb{R}$,
- la statistique doit être stochastiquement plus grande sous $\mathcal{H}_1$ que sous $\mathcal{H}_0$, i.e.
  $\mathbb{P}(w_i > t | i \in \mathcal{H}_1) > \mathbb{P}(w_i > t | i \in \mathcal{H}_0)$, pour tout $t \in \mathbb{R}$

On voit ici apparaître des conditions très semblables à l'hypothèse A1 de symétrie du bruit et à l'hypothèse de positivité des sources. Nous verrons effectivement dans le paragraphe 6.3.1 comment construire ces statistiques à partir de ces hypothèses, sans utiliser de contrefaçons.

Barber et Candès établissent alors une procédure de contrôle du FDR que nous reformulons ici dans notre contexte :

Proposition 4 (Procédure BC)
On note $\mathcal{W} = \{|w_j|,\ 1 \leq j \leq n\}$. *Si les* $\{w_j\}_{1\leq j\leq n}$ *sont distribuées de façon symétrique sous* $\mathcal{H}_0$, *et si leurs signes sont indépendants entre eux, alors pour un niveau nominal de contrôle donné* q, *un seuillage au niveau*

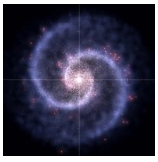
$$\widehat{t}_q = \min\left\{ t \in \mathcal{W} : \frac{1 + \#\{w_j \leq -t\}}{1 \vee \#\{w_j \geq t\}} \leq q \right\} \tag{6.13}$$

assure un contrôle exact du FDR au niveau q *pour l'ensemble de détections* $\mathcal{D} = \{i : w_i \geq \widehat{t}_q\}$.

La démonstration découle directement du Théorème 3 de [Barber et Candès 2015], que l'on applique ici à des p-valeurs binaires

$$p_{(i)} = \begin{cases} 1/2, & \text{si } w_{(i)} > 0, \\ 1, & \text{si } w_{(i)} < 0, \end{cases}$$

4. Notons qu'on considère ici les w_i comme des variables continues, on a donc pour tout t, $\mathbb{P}(w_i > t) = \mathbb{P}(w_i \geq t)$.



où $w_{(i)}$ sont les statistiques de contrôle ordonnées en valeur absolue i.e. $|w_{(1)}| \geq |w_{(2)}| \geq \cdots \geq |w_{(n)}|$.

Notons que trouver le seuil $\widehat{t}_q$ est équivalent à définir une procédure séquentielle sur les signes des statistiques de contrôle ordonnées en valeur absolue. En effet il vient que $\widehat{t}_q = w_{\widehat{k}}$ où $\widehat{k} = \max_k\{k : \frac{1+\#\{w_{(i)}<0,\, 1\leq i\leq k\}}{1\vee\#\{w_{(i)}>0,\, 1\leq i\leq k\}} \leq q\}$. Travailler sur les statistiques ordonnées permet ainsi de surveiller de manière séquentielle uniquement les signes, sans perte de puissance (voir [Barber et Candès 2015] pour plus de détails).

6.3.1 Construction de statistiques de contrôle

La construction de contrefaçons peut être problématique notamment en grande dimension. Afin d'éviter la construction de ces dernières, nous proposons ici de nous appuyer sur les statistiques de test construites dans le paragraphe 6.1. Chaque statistique est alors définie comme :

$$w_i = w_i^+ \vee w_i^- \times \begin{cases} +1 \text{ si } w_i^+ > w_i^-, \\ -1 \text{ si } w_i^+ < w_i^-, \end{cases} \tag{6.14}$$

où pour $1 \leq i \leq n$, $w_i^+ = \max_j\{S(\boldsymbol{d}_j, \boldsymbol{y}_i)\}$ et $w_i^- = -\min_j\{S(\boldsymbol{d}_j, \boldsymbol{y}_i)\}$, S étant la mesure de similarité (impaire, hyp. A2) choisie, et $\boldsymbol{d}_j$ les atomes du dictionnaire. Pour un bruit à distribution symétrique (hyp. A1), ces statistiques de contrôle sont symétriques sous $\mathcal{H}_0$ alors que de fortes valeurs positives sont attendues pour les pixels de la cible.

6.3.2 Relation avec la procédure BH empirique

Il est à noter que la procédure de contrôle par calibration des p-valeurs développée dans le paragraphe 6.1 peut s'exprimer de façon très proche de la procédure de contrôle de Barber et Candès pour ces statistiques de contrôle.

Rappelons en effet que p_i désigne la p-valeur associée à $i$eme statistique de test, $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n)}$, les p-valeurs ordonnées, et $\mathcal{H}_0^{(1)}, \ldots, \mathcal{H}_0^{(n)}$ les hypothèses nulles pour cet ordonnancement. Pour un contrôle fixé $0 \leq q \leq 1$, la procédure BH_q adaptative rejette $\mathcal{H}_0^{(1)}, \ldots, \mathcal{H}_0^{(\hat{k})}$ où

$$\hat{k} = \max\left\{0 \leq k \leq n :\ p_{(k)} \leq \frac{q}{\hat{\pi}_0}\frac{k}{n}\right\},$$

où $\hat{\pi}_0$ est un estimateur de π_0 (estimateur de Storey par exemple), et $p_{(0)} = 0$ par convention. Dans notre contexte de test unilatéral à droite, la $i$eme p-valeur *empirique*, est obtenue à partir de la distribution empirique sous $\mathcal{H}_0$ comme

$$p_i = 1 - \widehat{F}_0\left(w_i^+\right) = \widehat{G}_0(w_i^+), \quad \text{pour } 1 \leq i \leq n. \tag{6.15}$$

La distribution empirique sous l'hypothèse nulle $\widehat{F}_0$ est construite à partir des statistiques des minima. D'après (6.10) et (6.11), il vient que

$$\widehat{G}_0(t) = 1 - \widehat{F}_0(t) = \frac{1}{\hat{\pi}_0}\frac{\#\{w_i^- > t\}}{n} \approx \frac{1}{\hat{\pi}_0}\frac{\#\{w_i^- \geq t\}}{n},$$

pour tout $t \geq \widehat{\mu}_0$, et où $\hat{\pi}_0$ est équivalent à l'estimateur de Storey d'après la proposition 3.

Par conséquent, pour tout k tel que $w_k^+ \geq \widehat{\mu}_0$, la $k$ième p-valeur empirique (6.15) peut se réexprimer comme

$$p_k = \frac{1}{\hat{\pi}_0} \frac{\#\{w_i^- \geq w_k^+\}}{n}.$$

La procédure BH adaptative se réduit donc ici à

$$\begin{aligned} \hat{k} &= \max_k \left\{ \frac{1}{\hat{\pi}_0} \frac{\#\{w_i^- \geq w_{(k)}^+\}}{n} \leq \frac{q}{\hat{\pi}_0} \frac{k}{n} \right\}, \\ &= \max_k \left\{ \frac{\#\{w_i^- \geq w_{(k)}^+\}}{\#\{w_i^+ \geq w_{(k)}^+\}} \leq q \right\}, \end{aligned}$$

avec la convention $\hat{k} = 0$ si l'ensemble est vide, et où $w_{(1)}^+ \geq \ldots \geq w_{(n)}^+$. Si nous considérons des statistiques de contrôle construites à partir d'un seul atome, alors $w_i^+ = -w_i^- \equiv w_i$ et la procédure BH empirique est alors équivalente à trouver un seuil t à l'aide de la proposition 4. Il est à noter que, pour assurer un contrôle exact du FDR pour tout niveau de contrôle, la proposition 4 ajoute un 1 au numérateur (voir [Barber et Candès 2015]). Dans la procédure BH empirique cela se traduirait par une modification de l'estimation de quantiles afin de calculer les p-valeurs empiriques qui ne s'écrivent plus $p_k = \frac{1}{\hat{\pi}_0} \frac{\#\{w_i^- \geq w_k^+\}}{n}$ mais $p_k = \frac{1}{\hat{\pi}_0} \frac{1+\#\{w_i^- \geq w_k^+\}}{n}$.

L'avantage de la procédure BH empirique sous $\mathcal{H}_0$ est qu'il est possible d'améliorer l'estimation de la distribution empirique sous $\mathcal{H}_0$ en travaillant sur un plus grand nombre d'échantillons (par exemple une fenêtre spatiale plus grande que la zone de test peut être utilisée pour l'estimation). Mais cette procédure ne permet pas prendre en compte d'*a priori* de connexité comme nous allons désormais le faire en généralisant la procédure BC.

6.3.3 Algorithme

6.3.3.1 Algorithme générique

Considérons désormais un problème de détections multiples dans lequel les échantillons ciblés sont possiblement structurés (e.g. la cible est un objet connexe et large de plusieurs pixels dans une image). Nous proposons un algorithme générique prenant en compte cette connexité spatiale afin d'améliorer la puissance de détection du test (6.13), tout en assurant le même contrôle du FDR.

La stratégie proposée est la suivante : la cible étant possiblement formée des zones connexes, seuls les voisins des pixels déjà sélectionnés sont à tester et ainsi de suite jusqu'à atteindre un critère d'arrêt. Un exemple simple d'une telle stratégie consiste à développer une croissance de région. L'intérêt d'une telle stratégie est double : cela permet d'une part de favoriser des *a priori* de connexité sur la solution recherchée, et d'autre part de définir une procédure adaptative où le nombre de tests à effectuer réellement dépend des données. **La réduction du nombre effectif de tests aux seuls voisins pemet alors de limiter la perte de puissance pour un niveau de FDR donné.** Cette approche est décrite dans la procédure "step-up" de l'algorithme 6. Le contrôle du FDR s'appuie sur la propriété de symétrie de la distribution des statistiques de contrôle sous $\mathcal{H}_0$. La difficulté vient du biais de sélection qui est en général

introduit par la stratégie de recherche. Par exemple, si seules les plus grandes statistiques positives sont sélectionnées, on peut aisément vérifier que la distribution des statistiques n'est plus symétrique sous $\mathcal{H}_0$. Nous allons donc voir comment assurer l'absence de biais de sélection. Notons que par soucis de simplicité, chacun des n pixels est identifié par un indice $i \in \{1, \dots, n\}$ plutôt que par ses coordonnées spatiales. La connexité s'entend toutefois bien ici en terme de coordonnés spatiales. A une étape de sélection k donnée, notons $\mathcal{A}_k \subset \{1, \dots, n\}$ l'ensemble des pixels d'"intérêt" sélectionnés. Pour éviter tout biais de sélection, la procédure de sélection S_c de l'ensemble $\mathcal{A}_k$ doit préserver P1 :

P1 (Symétrie post-sélection)**.** *Pour tout $j \in \mathcal{A}_k$ correspondant à une vraie hypothèse nulle, w_j est distribué symétriquement.*

Cette procédure de sélection peut être définie de diverses façons, l'objectif étant de promouvoir la connexité spatiale avec les pixels nouvellement sélectionnés. Nous verrons dans le paragraphe 6.3.3.2 comment une telle procédure peut être construite.

Une estimation de la proportion de fausses découvertes, ou FDP, (parmi les statistiques positives $w_i > 0$ sélectionnés dans $\mathcal{A}_k$) est défini par

$$\widehat{q}_k = \frac{1 + \#\{i \in \mathcal{A}_k, w_i < 0\}}{1 \vee \#\{i \in \mathcal{A}_k, w_i > 0\}}. \tag{6.16}$$

L'algorithme construit donc itérativement un ensemble de test $\mathcal{A}_{\widehat{k}}$ (défini à l'étape 8 de l'algorithme 6) à l'aide d'une procédure de sélection S_c vérifiant la propriété P1, tout en contrôlant à chaque itération la valeur estimée du FDP. L'ensemble $\mathcal{D}$ des pixels constituant la cible est ensuite obtenu par une procédure de test sur les statistiques de contrôle de $\mathcal{A}_{\widehat{k}}$. Notons que cet algorithme nécessite pour initialisation que l'ensemble $\mathcal{A}_0$ soit non vide, autrement dit, on peut définir un ou plusieurs pixels de "départ" (dans notre application, il est aisé de considérer pour cela le centre de la galaxie).

Algorithme 6 Procédure COMET générique

1: *Entrée :* statistiques de contrôle $\boldsymbol{w} = \{w_j\}_{1 \le j \le n}$, niveau de contrôle nominal q

2: $k \leftarrow 0$, $\mathcal{A}_0$, $\widehat{q}_0 \leftarrow 0$ ▷ Initialisation
3: **while** $\mathcal{A}_k \neq \{1, \dots, n\}$ **do**
4: $\quad \mathcal{A}_{k+1} \leftarrow S_c(\mathcal{A}_k)$ ▷ Étape de sélection vérifiant P1
5: $\quad$ Calcul de $\widehat{q}_{k+1}$ à l'aide de (6.16) ▷ Estimation FDP
6: $\quad k \leftarrow k + 1$
7: **endWhile**
8: $\widehat{k} \leftarrow \max\{k : \widehat{q}_k \le q\}$ ▷ Temps d'arrêt
9: *Sortie :* $\mathcal{D} \leftarrow \{i \in \mathcal{A}_{\widehat{k}} : w_i > 0\}$ ▷ Liste des détections

Cet algorithme peut s'appuyer sur de nombreuses implémentations de la procédure de sélection sous réserve de préserver P1. Il s'agit principalement de trouver un bon ordonnancement des pixels à tester afin d'assurer la puissance de la procédure de détection. Pour favoriser la sélection des pixels de la cible cet ordonnancement doit ainsi d'une part privilégier les pixels/statistiques les moins vraisemblables sous $\mathcal{H}_0$ et d'autre part s'adapter aux pixels déjà sélectionnés afin de privilégier ici les a priori de connexité. Ce second point peut être vu comme la principale

différence avec la procédure BH (et l'approche de Barber et Candès) qui s'appuie sur un ordonnancement global des statistiques de contrôle par valeur absolue. Dans la suite nous nous focalisons sur une procédure de sélection simple qui donne de bons résultats expérimentaux.

6.3.3.2 Implémentation

On propose ici une méthode simple de croissance de région où le pixel du voisinage le moins vraisemblable selon $\mathcal{H}_0$ (au sens de la valeur absolue de la statistique w_i) est accrété à la région de pixels déjà sélectionnés. Plus précisément, à l'étape k, notons $\mathcal{N}_k = G(\mathcal{A}_k)$ le voisinage externe de $\mathcal{A}_k$, où G est le gradient morphologique externe, i.e. une dilatation (ici pour une clique en 8-connexité) suivie d'une soustraction. La procédure de sélection est alors définie par

$$S_c(\mathcal{A}_k) \equiv \mathcal{A}_k \cup \{j_0\}, \qquad \text{où } j_0 = \arg\max_{j \in \mathcal{N}_k} |w_j|.$$

Cette procédure est illustrée sur la figure 6.6.

La propriété de symétrie P1 est alors assurée par le lemme 2. Cette approche gloutonne permet de s'adapter à n'importe quelle forme connexe mais également de surmonter des "trous" puisqu'on cherche le plus grand ensemble de pixels où $\widehat{q} \leq q$.

Lemme 2
Si le vecteur $\boldsymbol{w} = (w_1, \ldots, w_n)$ est distribué de façon symétrique, et que l'opérateur de sélection S_c dépend des données uniquement via les valeurs absolues $|w_i|$ des statistiques de contrôle, pour $1 \leq i \leq n$, alors la propriété P1 est satisfaite pour tout ensemble $\mathcal{A}$ construit à l'aide de S_c.

Démonstration. Sous $\mathcal{H}_0$, par symétrie de la loi du vecteur $\boldsymbol{w}$, montrons tout d'abord que l'on a indépendance pour tout i entre $\mathrm{sgn}(w_i)$ et les $|w_j|, 1 \leq j \leq n$. En effet $\forall \epsilon \in \{+1, -1\}$, $t_j \geq 0$ pour $1 \leq j \leq n$, $\mathbb{P}(\epsilon w_i > t_i, |w_j| > t_j \; \forall j \neq i) = \mathbb{P}(\epsilon w_i < -t_i, |w_j| > t_j \; \forall j \neq i)$ (car $\boldsymbol{w}$ et $-\boldsymbol{w}$ ont la même distribution).

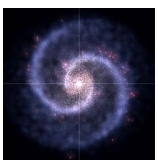
D'où

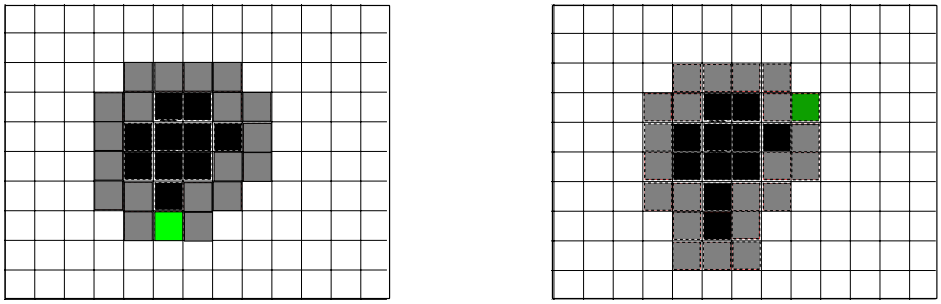
$$\begin{aligned}
\mathbb{P}\left(\mathrm{sgn}(w_i) = \epsilon, |w_j| > t_j, \; \forall j\right) &= \mathbb{P}\left(\epsilon w_i > t_i, |w_j| > t_j, \; \forall j \neq i\right) \\
&= \frac{1}{2}\mathbb{P}\left(|w_j| > t_j, \; \forall j\right) \\
&= \mathbb{P}\left(\mathrm{sgn}(w_i) = \epsilon\right) \times \mathbb{P}\left(|w_j| > t_j, \; \forall j\right).
\end{aligned}$$

On a donc bien indépendance entre $\mathrm{sgn}(w_i)$ et les $|w_j|, 1 \leq j \leq n$.

On note $\mathcal{A}$ l'ensemble des pixels sélectionnés par la procédure de sélection S_c à partir d'un ensemble de pixels dans lequel les w_i sont déjà distribués selon une loi symétrique sous $\mathcal{H}_0$. Pour $t \geq 0$, sous $\mathcal{H}_0$ on a alors,

$$\begin{aligned}
\mathbb{P}(w_i > t | i \in \mathcal{A}) &= \mathbb{P}(|w_i| > t, w_i > 0 | i \in \mathcal{A}) \\
&= \mathbb{P}(|w_i| > t, i \in \mathcal{A} | w_i > 0) \times \frac{\mathbb{P}(w_i > 0 | i \in \mathcal{A})}{\mathbb{P}(i \in \mathcal{A} | w_i > 0)} \\
&= \mathbb{P}(|w_i| > t, i \in \mathcal{A}) \times \frac{\mathbb{P}(w_i > 0)}{\mathbb{P}(i \in \mathcal{A})}
\end{aligned}$$





(a) Etape k (b) Etape $k+1$

FIGURE 6.6 – Procédure de sélection : à chaque étape, on explore le voisinage $\mathcal{N}_k$ (en gris) des pixels déjà sélectionnés $\mathcal{A}_k$ (en noir) et on choisit le pixel avec la plus grande statistique en valeur absolu (j_0, en vert). On met alors à jour la zone sélectionnée et le voisinage et on itère le processus.

La première égalité est immédiate, la deuxième égalité provient des formulations des lois conditionnelles et la troisième égalité vient du fait que les événements $|w_i| > t$ et $i \in \mathcal{A}$ sont indépendants du signe de w_i. De façon similaire

$$\mathbb{P}(-w_i > t | i \in \mathcal{A}) = \mathbb{P}(|w_i| > t, i \in \mathcal{A}) \times \frac{\mathbb{P}(w_i < 0)}{\mathbb{P}(i \in \mathcal{A})}$$

Sous $\mathcal{H}_0$, $\mathbb{P}(w_i < 0) = \mathbb{P}(w_i > 0)$ d'où $\mathbb{P}(w_i > t | i \in \mathcal{A}) = \mathbb{P}(-w_i > t | i \in \mathcal{A})$ ce qui conclut la preuve. □

La condition de symétrie du vecteur $\boldsymbol{w} = [w_1, \ldots, w_n]$ se traduit par la nécessité (du fait de l'hypothèse A2 d'imparité de la fonction de construction des statistiques de test) pour la distribution jointe des vecteurs de bruit $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n$ d'être symétrique, i.e. $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n$ et $-\boldsymbol{\epsilon}_1, \ldots, -\boldsymbol{\epsilon}_n$ ont la même distribution jointe. Notons que bien que la symétrie des lois marginales des $\boldsymbol{\epsilon}_i$ n'implique pas nécessairement la symétrie de la loi jointe, cette équivalence est en pratique vérifiée pour la plupart des distributions classiques, comme les familles elliptiques.

Notons qu'en pratique, pour des gains de temps de calcul, la boucle interne de l'algorithme 6 peut être arrêtée lorsqu'à la fois le nombre de pixels sélectionnés est grand et $\widehat{q}_k$ est significativement plus grand que q (e.g. $\widehat{q}_k \geq 1.2 \times q$).

Remarque 3
Il est à noter que cette notion de FDR "connexe" peut être assimilée à la notion de pureté fréquemment utilisée en astronomie (voir par exemple SERRA *et al. 2012 ;* CHIU *et al. 2016).*

6.3.4 Résultats théoriques

Proposition 5 (Contrôle FDR de COMET)
Supposons que les vecteurs de bruit $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n$ sont distribués selon une loi symétrique et sont indépendants entre eux. Alors l' algorithme 6, où les statistiques de contrôle $\boldsymbol{w}$ sont construites à l'aide de (6.14), permet un contrôle exact du FDR : $\mathbb{E}\left[\frac{U}{R \vee 1}\right] \leq q$.

Démonstration. La propriété P1 est assurée par le lemme 2. Pour tout $i \in \mathcal{A}_{\widehat{k}}$ correspondant à un vrai $\mathcal{H}_0$, w_i est donc symétriquement distribué. De plus les signes des $\{w_i\}_{1\leq i\leq n}$ sont indépendants par indépendance des $\{\boldsymbol{\epsilon}_i\}_{1\leq i\leq n}$. Nous pouvons donc appliquer la proposition 4 avec un seuil $\widehat{t}_q = 0$. Cela conclut la preuve. □

En cas de bruit corrélé, le contrôle exact n'est plus assuré. Néanmoins un contrôle asymptotique peut être prouvé à l'aide des hypothèses suivantes.

A6 (Faible dépendance)**.** *Pour n'importe quel ensemble $\mathscr{S}$ de pixels sous $\mathcal{H}_0$,*

$$\frac{\#\{i \in \mathscr{S} : w_i > 0\}}{\#\mathscr{S}} \underset{\#\mathscr{S}\to\infty}{\overset{a.s.}{\to}} \mathbb{P}(w_i > 0)$$

Pour simplifier les notations, $\mathscr{S}_n$ indique désormais l'ensemble final de pixels sélectionnés $\mathcal{A}_{\hat{k}}$ par l'algorithme 6 pour un jeu de données de n pixels.

A7 (Croissance de région)**.** *Pour un niveau de contrôle $q > 0$ donné, $\mathscr{S}_n$ croît avec le nombre total d'échantillons n et $\#\mathscr{S}_n \underset{n\to+\infty}{\to} +\infty$.*

L'hypothèse A6 implique que les corrélations entre pixels sont de faible portée. C'est bien le cas sur les données MUSE (avant prétraitement de type filtrage adapté spatial, voir paragraphe 6.3.6) car les corrélations proviennent essentiellement des étapes de *drizzling* (voir chapitre 1). L'hypothèse A7 signifie simplement que pour atteindre le régime asymptotique, il faut que la taille de la source grandisse avec le nombre d'échantillons. Autrement dit (et contrairement à l'approche BH empirique où on peut apprendre la loi de $\mathcal{H}_0$ sur un nombre d'échantillons arbitrairement grand), il ne suffit d'explorer une région de plus en plus grande si π_0 tend vers 1.

Proposition 6 (Contrôle asymptotique de COMET)
Supposons A6, A7 et une distribution symétrique des statistiques de contrôle sous $\mathcal{H}_0$. Alors l'algorithme 6 permet un contrôle asymptotique du FDR pour le test (6.1).

Démonstration. Notons $m = \#\mathscr{S}_n$, $m^+ = \#\{i \in \mathscr{S}_n, w_i > 0\}$, $m^- = \#\{i \in \mathscr{S}_n, w_i < 0\}$, $m_0^- = \#\{i \in \mathscr{S}_n \cap \mathcal{H}_0, w_i < 0\}$ et $m_0^+ = \#\{i \in \mathscr{S}_n \cap \mathcal{H}_0, w_i > 0\}$. Alors

$$\widehat{\mathrm{FDP}}_q = \frac{1 + m^-}{m^+}$$

et

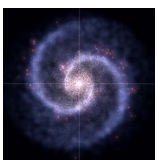
$$\mathrm{FDP} = \frac{m_0^+}{m^+}.$$

Comme $\widehat{\mathrm{FDP}}_q \geq \frac{m_0^-}{m^+}$, nous avons ($m^- \geq m_0^-$ par construction) :

$$\widehat{\mathrm{FDP}}_q - \mathrm{FDP} \geq \frac{m_0^- - m_0^+}{m^+}.$$

De plus, d'après l'hypothèse de symétrie sous $\mathcal{H}_0$, $\mathbb{P}(i \in \mathcal{H}_0, w_i > 0) = \mathbb{P}(i \in \mathcal{H}_0, w_i < 0)$. Par application des hypothèses A6 et A7, on en déduit donc que

$$\liminf_{n\to\infty} \frac{m_0^-/m - m_0^+/m}{m^+/m} = 0.$$



D'où $\liminf_{n\to\infty}(\widehat{\mathrm{FDP}}_q - \mathrm{FDP}) \geq 0$. Puisque $\widehat{\mathrm{FDP}}_q \leq q$ par construction, on obtient que

$$\limsup_{n\to\infty} \mathrm{FDP} \leq q$$

avec une probabilité de 1. Alors, par application du lemme de Fatou [5],

$$\limsup_{n\to\infty} \underbrace{\mathbb{E}[\mathrm{FDP}]}_{\mathrm{FDR}} \leq q,$$

ce qui conclut la preuve. □

6.3.5 Validation sur simulation

Nous comparons ici la procédure COMET avec la méthode décrite en proposition 4, page 101, qui repose sur une procédure BH empirique mais ne prend pas en compte d'information de connexité. Les données simulées sont construites sous forme de cubes hyperspectraux de dimension $31 \times 51 \times 51$ (une dimension spectrale et 2 spatiales). La cible est composée de 350 pixels connexes associés à un spectre. Chaque spectre est de forme gaussienne tronquée, centrée sur la bande spectrale médiane $j = 15$. Du bruit gaussien est ajouté, corrélé par convolution d'un noyau de taille 3×3. Le pRSB est d'environ 7dB (le pRSB est défini comme $\mathrm{RSB} = 20 \log \frac{u_m}{\sigma}$, où u_m est la valeur maximale de la cible et σ est l'écart-type du bruit, une source réelle MUSE ayant un pRSB typiquement compris entre 0dB et 10dB). Pour simplifier les simulations, les statistiques sont construites à partir d'un seul atome : pour une signature de cible $\boldsymbol{d}$ connue, la statistique w_i du $i$eme test, $1 \leq i \leq n$, est définie par :

$$w_i = \boldsymbol{d}^T \boldsymbol{y}_i \tag{6.17}$$

qui correspond au filtre adapté dans le cas d'un bruit blanc. Rappelons (voir paragraphe 6.3.2 page 102) que dans ce cadre (un seul atome), l'approche BC définie dans la proposition 4 est similaire à l'approche BH empirique développée dans le paragraphe 6.2.

Comme décrit dans la table 6.2, prendre en compte la connexité améliore drastiquement la puissance de détection tout en assurant un contrôle du FDR. De plus nos résultats illustrent la robustesse de COMET vis à vis de la taille de la région testée : la puissance de détection reste sensiblement constante. La figure 6.7 illustre la procédure de détection : la figure 6.7(a) représente une réalisation bruitée des données testées, intégrée le long de la dimension spectrale ; les figures 6.7(b) et 6.7(c) montrent les taux de détections sur 100 simulations Monte-Carlo pour respectivement l'approche non-connexe et COMET. Le code couleur qui représente le taux de détection illustre le gain de puissance obtenu avec COMET. La région explorée par COMET (contour blanc) reste à proximité de la cible, au contraire de la méthode non-connexe (presque tous les pixels ont été détectés au moins une fois sur les 100 simulations dans la fig. 6.7(b)). La figure 6.8 détaille les performances de la méthode pour différents niveaux de FDR : la figure 6.8(a) souligne l'obtention d'un contrôle asymptotique en présence de bruit faiblement corrélé, pour les deux approches ; la figure 6.8(b) illustre à nouveau le gain significatif de puissance lors

5. Si $f_1, f_2, \ldots$ est une suite de fonctions mesurables sur un espace mesuré X, à valeurs positives, alors

$$\int_X \liminf_{n\to\infty} f_n \leq \liminf_{n\to\infty} \int_X f_n$$

	COMET	Méthode non-connexe
Région 51×51		
FDR (%)	4.85	4.74
Puissance (%)	76.1	35.4
Région 71×71		
FDR (%)	4.84	4.65
Puissance (%)	76.0	25.5

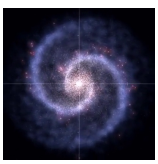
TABLE 6.2 – Comparaison de la puissance entre COMET et la procédure BH empirique, définie dans le paragraphe 6.2, sans *a priori* de connexité. Le FDR nominal est de 5%. La cible fait 350 pixels, sur un total de $51 \times 51 = 2601$ pixels dans le premier cas et de $71 \times 71 = 5041$ dans le second. Les résultats sont moyennées sur 400 simulations Monte-Carlo.

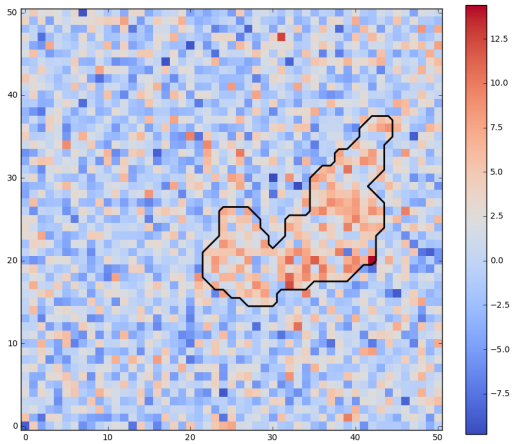
de la prise en compte de la connexité. De plus, COMET garde la même puissance de détection lorsque la fenêtre de tests est agrandie alors que l'approche classique voit sa puissance diminuer (rappelons en effet que dans la procédure BH on pénalise les p-valeurs en fonction du nombre de tests effectués).

6.3.6 Impact des corrélations spatiales

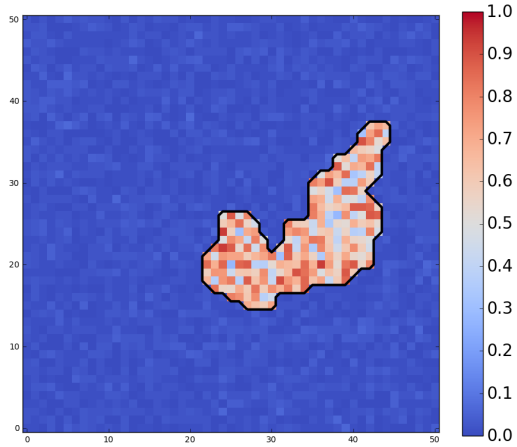
Du fait du faible RSB, il est assez naturel de chercher à améliorer la puissance à l'aide d'un filtrage adapté, en convoluant par la FSF ou simplement par un noyau gaussien. Toutefois, cela n'est pas sans conséquence sur l'analyse qui en suit. On peut notamment remarquer que l'utilisation d'un filtre gaussien conjointement à l'analyse de courbes de niveaux, méthode fréquemment utilisée en astronomie, n'est pas sans risque. En effet on peut voir sur la figure 6.9 que sur données simulées, les courbes de niveaux semblent rapidement indiquer des extensions rattachées à la galaxie bien en dehors du signal simulé dès lors que la taille du noyau appliqué devient grande. Notons en effet que l'image 6.9(c) est obtenue pour un noyau semblable à la FSF caractéristique de MUSE. On voit donc qu'il peut être fallacieux de seuiller directement à partir des courbes de niveaux obtenues dès lors qu'un filtre spatial adapté à la FSF a été appliqué en prétraitement.

Notons qu'en présence de fortes corrélations, COMET ne contrôle pas non plus le FDR car les tailles des sources ne sont pas suffisamment grandes devant la corrélation spatiale du bruit pour que les hypothèses asymptotiques requises par le théorème 6 soient valides. En effet, comme on peut voir sur la figure 6.10, plus les corrélations sont importantes pour une source de taille fixe, plus on s'éloigne du régime asymptotique dans lequel le contrôle du FDR est effectif. La figure 6.10 permet également d'illustrer le fait que la procédure BH empirique, moins puissante, peut éviter cet écueil grâce à un apprentissage de la distribution du bruit sur une région suffisamment large, sous l'hypothèse de stationnarité du bruit. Du fait de la grande puissance de détection de COMET, on peut toutefois se contenter d'un filtrage par un petit noyau spatial (vis à vis de la taille de la source) ce qui assure une bonne puissance de détection avec un contrôle préservé. En pratique on se limitera sur les données réelles à des noyaux 3 par 3 obtenus en tronquant la FSF, ce qui permet dans les données MUSE de s'assurer que la distance caractéristique des corrélations reste petite devant la taille caractéristique d'une source.

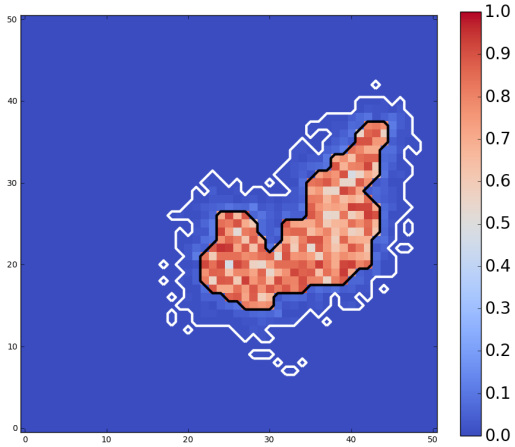




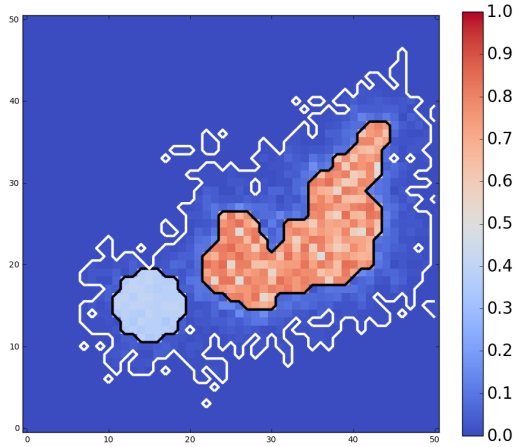
(a) Réalisation bruitée



(b) Détection par méthode non-connexe

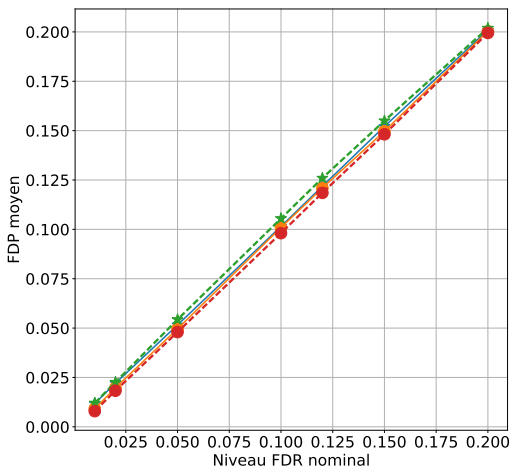


(c) Détection par COMET

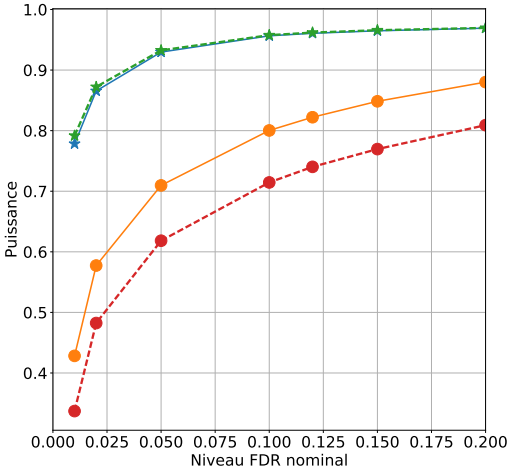


(d) Détection par COMET sur une cible non-connexe

FIGURE 6.7 – (a) exemple d'une réalisation bruitée de cube simulé (intégré sur la dimension spectrale) ; (b) taux de détection avec un FDR de 10% pour la méthode de la prop. 4 (1 : toujours détecté, 0 : jamais détecté) ; (c) taux de détection par COMET pour un même niveau FDR de 10% ; (d) taux de détection de COMET sur une cible non-connexe. Résultats moyennés sur 100 simulations Monte-Carlo. En noir : position de la cible, en blanc : pixels détectés par COMET au moins une fois parmi les 100 simulations.

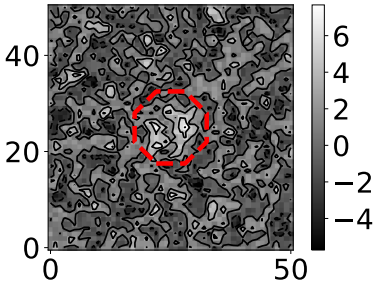


(a) Contrôle FDR : FDR empirique vs FDR nominal

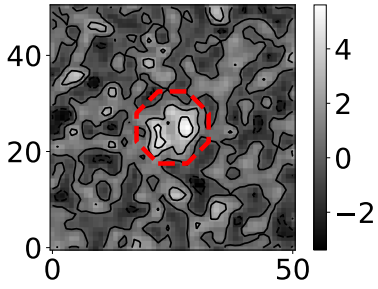


(b) Puissance de détection : puissance vs FDR nominal

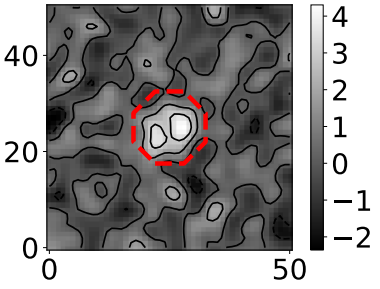
FIGURE 6.8 – Méthode proposée : $\star$, l'approche non-connexe : $\bullet$. En trait plein, la détection est faite sur une fenêtre 51×51, en pointillés sur une zone 71×71 (même cible de taille 360 pixels). Résultats moyennés sur 500 simulations Monte-Carlo.



(a) Courbes de niveaux pour $\sigma = 0.2$

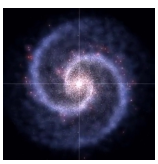


(b) Courbes de niveaux pour $\sigma = 0.8$



(c) Courbes de niveaux pour $\sigma = 1.4$

FIGURE 6.9 – Evolution des courbes de niveaux sur l'image bande étroite d'une source simulée (de taille 185 pixels au sein d'une région de 51 par 51 pixels) en fonction de la taille σ du noyau gaussien de convolution appliqué en prétraitement. En pointillés, le support de la source.



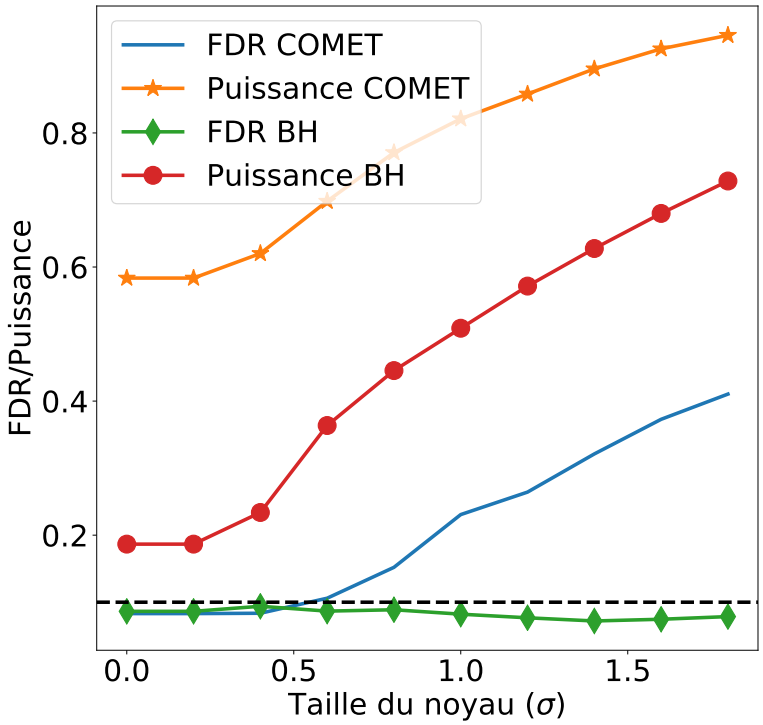


FIGURE 6.10 – Evolution de la puissance et du FDR sur une cible simulée en fonction de la taille σ du noyau gaussien de convolution appliqué en prétraitement. Résultats moyennés sur 20 tirages Monte-Carlo. La puissance de COMET est indiquée en ⋆, et son FDR par un trait plein bleu. La puissance de la procédure BH est en ● et son FDR en ♦. Le niveau de contrôle nominal, fixé à 0.1 est indiqué en pointillés noirs. La cible est de taille 185 pixels, recherchée sur une région de taille 51 par 51 pixels (voir fig. 6.9). La procédure BH empirique apprend la loi du bruit sur une région 101 par 101 pixels.

6.4 Construction du dictionnaire

Une des principales caractéristiques de la cible recherchée est que sa signature spectrale peut varier, cette variabilité étant en première approximation assimilée à une translation spectrale. Le dictionnaire est donc construit ici en créant des versions translatées d'une signature de référence, $\boldsymbol{d}_*$. En supposant que $\boldsymbol{d}_*$ provient de l'échantillonnage d'un modèle continu $f(\cdot)$, nous pouvons définir $\boldsymbol{d}_*^\delta$ les vecteurs des translatées obtenues en échantillonnant $f(\cdot - \delta)$. Le modèle de dictionnaire par *translatées linéairement espacées* (TLE) sur un intervalle $[-\tau, \tau]$ est alors défini, pour une taille m donnée, comme le dictionnaire $\boldsymbol{D}^m$ composé des atomes $\boldsymbol{d}_k = \boldsymbol{d}_*^{\tau_k}$, où $\tau_k = -\tau + \frac{2\tau}{m-1}k$, pour $k = 0, \ldots, m-1$.

La question clé est alors le choix du nombre m de versions translatées, ou, en d'autres termes, la redondance du dictionnaire. Afin d'étudier le choix de ce paramètre, nous nous plaçons dans un contexte simplifié :

- le bruit est supposé i.i.d. et de loi $\mathcal{N}(0,1)$;
- la mesure de similarité est un filtre adapté spectral entre un atome du dictionnaire et le spectre testé, tel que défini en (6.4) ;
- le spectre de référence $\boldsymbol{d}_*$ est un vecteur non-négatif de ℓ_2-norme unitaire, avec une fonction d'autocorrélation $\Gamma(u) = \langle \boldsymbol{d}_*, \boldsymbol{d}_*^u \rangle$ décroissante en $|u|$, et un support compact tel que $||\boldsymbol{d}_*^u|| = ||\boldsymbol{d}_*|| = 1$, pour $u \in [-\tau, \tau]$;
- la signature de la source $\boldsymbol{x}$ est construite à partir d'une translation $\boldsymbol{d}_0^u$ du spectre de référence $\boldsymbol{x} = a\boldsymbol{d}_0^u$, où $a > 0$, et u est une translation aléatoire qui est uniformément distribuée sur $[-\tau, \tau]$.

Une mesure de la redondance d'un dictionnaire normalisé $\boldsymbol{D}$ peut être donnée par sa cohérence, qui est définie par $\mu = \max_{i \neq j} |\langle \boldsymbol{d}_i, \boldsymbol{d}_j \rangle|$. Pour un dictionnaire TLE $\boldsymbol{D}^m$, et sous les hypothèses susmentionnées, cette cohérence se réduit à la corrélation entre deux atomes consécutifs : $\mu = \langle \boldsymbol{d}_j, \boldsymbol{d}_{j+1} \rangle$, pour $1 \leq j < m$. Comme illustré sur la figure 6.11, par construction du dictionnaire, plus la taille m du dictionnaire croît, plus les atomes sont corrélés et plus le dictionnaire est cohérent.

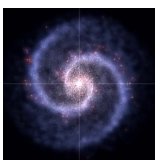
Notons $\boldsymbol{z}^m = (\boldsymbol{D}^m)^T \boldsymbol{y} \in \mathbb{R}^m$ le vecteur des statistiques du filtre adapté, dont les éléments sont définis comme z_j^m pour $1 \leq j \leq m$. Pour un seuil de décision donné η, la PFA pour l'approche max-test s'écrit

$$\alpha_m = \Pr\left(\max \boldsymbol{z}^m > \eta\right), \qquad \text{sous } \mathcal{H}_0. \tag{6.18}$$

Ici le vecteur de bruit $\boldsymbol{\epsilon}$ est distribué selon $\mathcal{N}(0, \boldsymbol{I}_m)$ sous $\mathcal{H}_0$. Si les atomes sont orthogonaux (par exemple, s'ils ont des supports disjoints), le vecteur $\boldsymbol{z}^m$ est alors distribué selon une loi normale de moyenne nulle et de matrice de covariance $\boldsymbol{D}^m(\boldsymbol{D}^m)^T = \boldsymbol{I}_m$. Dans ce cas, on peut calculer exactement la PFA par

$$\begin{aligned} \alpha_m &= 1 - \Pr\left(\max \boldsymbol{z}^m \leq \eta\right) = 1 - \Pr\left(z_1^m \leq \eta\right)^m, \\ &= 1 - \Phi\left(\eta\right)^m, \end{aligned} \tag{6.19}$$

où Φ est la fonction de répartition de la distribution normale. En pratique, le dictionnaire est choisi pour être fortement cohérent (car nous voulons suivre des translatées très proche du spectre de référence). Cela nécessite de trouver une autre façon d'estimer ou de borner cette probabilité.



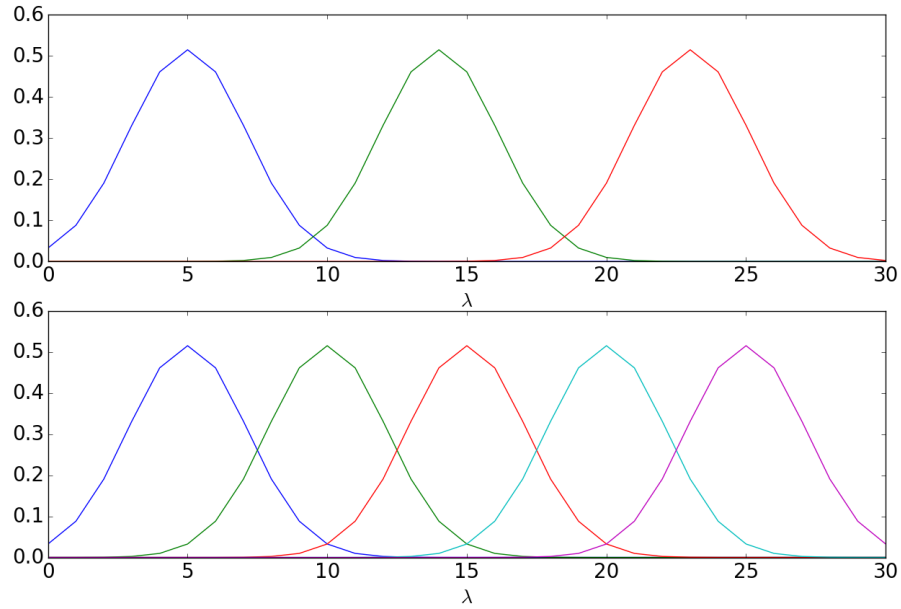


FIGURE 6.11 – Exemples de dictionnaires construits à partir d'une référence $\boldsymbol{d}_*$, avec un nombre variable d'atomes (3 atomes et une cohérence de 0.2 puis 5 atomes et une cohérence de 0.5). La référence $\boldsymbol{d}_* \in \mathbb{R}^l$, avec $l = 30$, est échantillonnée pour $j = 1, \ldots, l$ à partir d'une densité gaussienne centrée sur la bande médiane $j = 15$, avec une largeur à mi-hauteur de 5 échantillons ($\sigma \approx 2.12$) tronquée à ± 6 autour du mode, et ℓ_2-normalisée. Le décalage maximal est de $\tau = 8$ échantillons.

Proposition 7
Pour tout $t \in \mathbb{R}$ et $m \geq 2$, on définit $M_{m+1}(t)$ de façon récursive sous $\mathcal{H}_0$ par

$$M_{m+1}(t) = \Pr\left(z_1^{m+1} \leq t \mid z_2^{m+1} \leq t, z_3^{m+1} \leq t\right) \times M_m(t), \tag{6.20}$$

avec $M_2(t) = \Pr\left(z_1^2 \leq t, z_2^2 \leq t\right)$. Sous les hypothèses mentionnées précédemment, une borne supérieure de la PFA α_m est donnée par $1 - M_m(\eta)$.

Démonstration. Voir annexe A. □

L'intérêt de l'expression (6.20) est que le premier terme de droite et la valeur initiale $M_2(t)$ peuvent être évalués numériquement à l'aide de règles quadratiques pour des fonctions de distributions normales bivariées et trivariées (voir GENZ et BRETZ 2009), sans nécessité d'approximation par Monte-Carlo. Ainsi, cette borne supérieure peut être aisément calculée par récurrence avec précision. Quand les atomes sont décorrélés, cette borne est atteinte et se ramène à (6.19). De plus, elle permet de mesurer, pour un seuil η donné, l'évolution de la PFA α_m comme une fonction de la taille m du dictionnaire dans le cadre d'atomes fortement corrélés. De façon réciproque, cela permet d'évaluer le seuil η_m, qui assure un taux de fausses alarmes inférieur à un α donné pour tout $m \geq 1$.

Nous pouvons désormais estimer grossièrement le gain potentiel de détection sous $\mathcal{H}_1$ en fonction de m, de la façon suivante. Sous $\mathcal{H}_1$, nous avons supposé que

$$\boldsymbol{y} = a\boldsymbol{d}_*^u + \boldsymbol{\epsilon}$$

avec un décalage $u \sim \mathcal{U}([-\tau, \tau])$. Alors, si on suppose que le maximum est obtenu par l'atome le plus proche (au sens du décalage u), qui est par hypothèse le plus corrélé avec $\boldsymbol{d}_*^u$, la statistique de max-test moyenne peut être approximée par

$$E[\max \boldsymbol{z}^m] \approx aE[\Gamma(e_m)],$$

où $\Gamma(\cdot)$ est la fonction d'autocorrélation de $\boldsymbol{d}_*$ et $e_m \sim \mathcal{U}([0, \tau/(m-1)])$ est le décalage entre $\boldsymbol{d}_*^u$ et l'atome le plus proche.

A l'aide de cette valeur de max-test sous $\mathcal{H}_1$ et de la borne supérieure sur la fausse alarme donnée dans la proposition 7, nous pouvons voir dans la figure 6.12 que lorsque la taille m du dictionnaire croît, la statistique du max-test continue à croître sous $\mathcal{H}_1$. Toutefois, pour un niveau fixé de contrôle α, le seuil du test (ou sa borne supérieure) η_m ne croît pas de façon significative à partir d'une certaine taille, typiquement $m \geq 10$ dans la figure 6.12. Cela est clairement dû à l'accroissement des intercorrélations entre atomes lorsque m augmente. Au contraire, si les atomes sont décorrélés, nous pouvons voir que le seuil obtenu par l'équation (6.19) croît plus rapidement que le gain potentiel de la statistique du max-test sous $\mathcal{H}_1$.

Cela est confirmé par la figure 6.13, qui montre plusieurs courbes ROC empiriques pour différentes tailles m de dictionnaires. On peut voir empiriquement que plus le dictionnaire est cohérent, plus le max-test devient puissant.

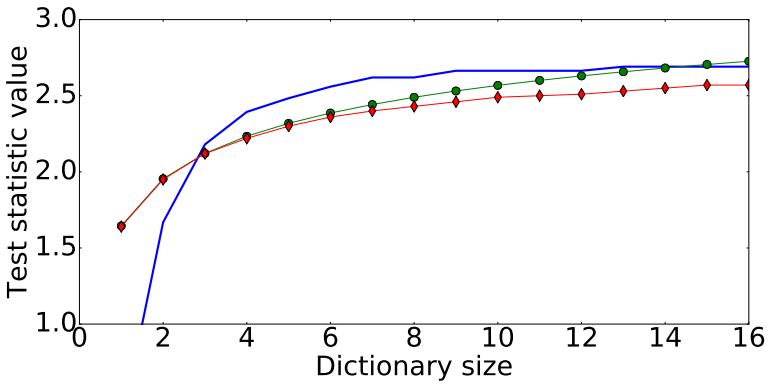
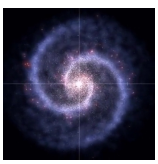


FIGURE 6.12 – Seuil des statistiques de test en fonction du nombre m d'atomes pour un niveau de contrôle PFA de $\alpha = 0.05$. Le seuil η_m pour le dictionnaire TLE $\boldsymbol{D}^m$ identique à la Fig. 6.11 ($m = 15$) est indiqué en ♦ ; le seuil pour un dictionnaire de taille m décorrélé est indiqué par • ; l'évolution du gain potentiel du test sous $\mathcal{H}_1$ pour une intensité $a = 2.7$ est représentée par en traits épais bleus.

En conséquence, dans notre application, le dictionnaire peut être construit pour être aussi cohérent que possible (jusqu'à attendre la fréquence d'échantillonnage spectrale de MUSE comme décalage spectral minimal). Dans le cadre de notre application de détection du halo entourant la galaxie, l'atome de référence $\boldsymbol{d}_*$ peut être estimé en moyennant spatialement les pixels centraux de la galaxie étudiée. Le spectre est limité à une bande spectrale de largeur $l = 30$ centrée autour du pic d'émission, ce qui suffit à assurer la présence entière de la raie d'émission. En s'appuyant sur des *a priori* astrophysiques, le décalage spectral est limité à l'intervalle $[-\tau, \tau]$ avec $\tau = 7$ bandes spectrales MUSE (i.e., $\tau \approx 9$ Å). Le décalage se fait à la résolution spectrale de MUSE pour éviter toute interpolation. Le dictionnaire $\boldsymbol{D}^m$ est finalement construit à partir des atomes correspondant à ces $m = 15$ translatées de $\boldsymbol{d}_*$.



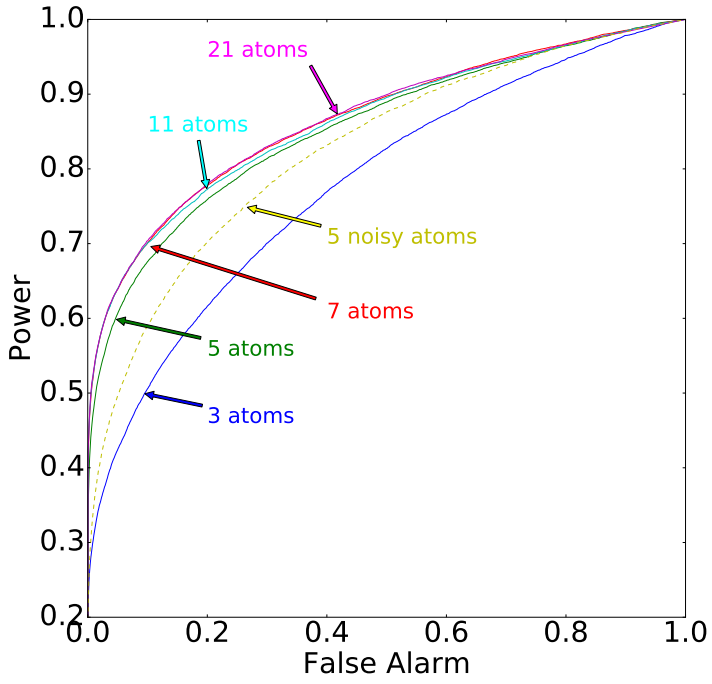


FIGURE 6.13 – Comparaison des courbes ROC empiriques pour plusieurs tailles m de dictionnaire sous le modèle de dictionnaire TLE. Les résultats sont obtenus sur 50 tirages Monte-Carlo de données simulées proches de MUSE, avec les hypothèses décrites dans le paragraphe 6.4.

Résumé

Dans ce chapitre nous avons formalisé le problème de détection du CGM comme un test d'hypothèses. Nous nous sommes appuyé sur les approches de type max-test sur des représentations 1-parcimonieuses des spectres explorés pour définir ce test. Nous avons mis en place une méthode originale et robuste de calibration de la statistique du bruit en s'appuyant sur la simple hypothèse de symétrie de la distribution du bruit. Cela nous permet d'estimer des p-valeurs et d'effectuer une détection sous contrôle du FDR à l'aide de procédure de contrôle classique BH. En regard de l'état de l'art, cette approche garantit une robustesse face aux erreurs de modélisation de la distribution du bruit tout en préservant la puissance de détection par rapport à des approches standards dans le cas gaussien. Toutefois, cette procédure ne permet pas de prendre en compte les structures spatiales d'une cible étendue comme le CGM. Nous avons alors développé une nouvelle méthode de contrôle du FDR, s'inspirant de récents travaux de Barber et Candès, permettant de prendre en compte une information de connexité spatiale. Cette approche permet d'améliorer de façon significative la puissance de détection du test, tout en conservant un contrôle robuste du FDR. Elle évite également une perte de puissance lorsque la taille de la fenêtre étudiée augmente. Nous avons également décrit la construction du dictionnaire de représentation des spectres dans le cadre de notre application.
Dans le chapitre suivant, nous allons désormais appliquer ces méthodes sur les données MUSE du champ profond UDF.

Chapitre 7

Application aux données MUSE

Sommaire

Dans ce chapitre, les méthodes développées dans le chapitre précédent sont appliquées sur les données UDF-10 de MUSE. Après un bref descriptif des données utilisées, les pré-traitements employés sont exposés dans le paragraphe 7.2 et les résultats sont présentés dans le paragraphe 7.3.

7.1 Données

Les données utilisées ici sont issues du champ UDF-10 observé par MUSE. Rappelons que le processus de réduction de données ainsi que la construction du catalogue de sources ponctuelles (galaxies) est décrit dans Bacon et al. 2017. Une dizaine de sources sont sélectionnées avec les experts, certaines semblant présenter un halo, d'autres non. Pour chacune de ces sources, on définit un voisinage spatio-spectral, centré spatialement sur le centre de la galaxie et spectralement sur le pic de la raie d'émission. On extrait ainsi pour chaque source un sous-cube de taille 51 par 51 pixels par 31 feuillets spectraux. Rappelons en effet que le signal d'intérêt (la raie d'émission Lyman-α) se concentre sur quelques feuillets spectraux autour de son pic.

L'objectif ici est donc d'explorer ce voisinage et de poursuivre la raie Lyman-α le plus loin possible de la galaxie. On suppose connu le centre spatial de la galaxie et la position spectrale de la raie d'émission dans le spectre de la galaxie. On suppose également que la raie émise par le halo a une forme similaire à la raie au sein de la galaxie, éventuellement déformée en première approximation par une translation spectrale. On construit donc un dictionnaire comme décrit dans le paragraphe 6.4. On suppose enfin que les pixels sont faiblement dépendants. Cette dernière hypothèse est justifiée car les dépendances entre pixels sont principalement dues aux interpolations lors des opérations d'alignement par *drizzling* (voir chapitre 1). On vérifie donc bien l'hypothèse A5. Notons que l'usage éventuel d'un filtrage par la FSF en prétraitement accentue significativement ces dépendances même si elles restent de courte portée, et biaise l'estimation du support spatial de la zone d'intérêt.

Les autres hypothèses requises pour appliquer les méthodes proposées dans le chapitre 6 sont également vérifiées. Les données MUSE résultent de la somme d'un grand nombre de poses

assurant la quasi-gaussianité (et donc la symétrie) du bruit, par application du théorème central limite. Les étapes de prétraitements (soustraction du fond, centrage, réduction de la variance, ...) conservent toutes la symétrie du bruit centré, assurant ainsi la validité de l'hypothèse A1. L'hypothèse A2 (mesure de similarité impaire) est vérifiée par construction (voir le choix de la mesure proposé au paragraphe suivant). Comme indiqué dans la remarque 2 du paragraphe 6.2, les hypothèses A3 et A4 ne peuvent être strictement garanties. Toutefois, en dehors de ce cadre idéal, le point clé est que l'équation (6.7) est une approximation suffisamment bonne. En effet, le signal ciblé est supposé suffisamment distinct du bruit de fond et bien approché par le dictionnaire (voir paragraphe 6.4) ; ainsi $T_{\max}(\boldsymbol{y})$ sera significativement plus grande sous $\mathcal{H}_1$ que sous $\mathcal{H}_0$ pour les signaux détectables. De plus les pixels de la galaxie et du halo sont supposés en forte minorité dans la région explorée spatialement, ce qui signifie que π_0 est proche de un, et améliore la qualité des approximations. Enfin, comme souligné dans la *remarque 2* (page 91), les erreurs d'approximation dans l'équation (6.7) ne peuvent entraîner qu'une faible perte de puissance mais le contrôle est toujours garanti (le biais est conservatif comme montré sur la figure 6.3 à faible RSB).

7.1.1 Mesure de similarité

La mesure de similarité retenue pour l'application aux données MUSE est la mesure SAD définie par (6.5). D'autres métriques ont été étudiées comme le filtre adapté (6.4) et la divergence spectrale (SID) [Chang 1999], basée sur la divergence de Kullback-Leibler symétrisée. Il s'agit d'une mesure bien adaptée à l'étude de spectre mais elle demande des signaux positifs ce qui ne peut pas être garanti dans notre cas du fait des nombreux pré-traitements et du faible RSB. Le filtre adapté donne des résultats satisfaisants mais le SAD semble plus robuste face à certaines erreurs systématiques et en présence de dynamiques d'intensité importantes. Par ailleurs, malgré la normalisation effectuée, la puissance du test utilisant le SAD s'avère sur des données types très proche de celle obtenue par filtre adapté.

7.2 Prétraitements

Afin de pouvoir appliquer notre méthode de détection sur les données MUSE, il est nécessaire de réaliser plusieurs pré-traitements. Ces prétraitements visent notamment à :

- augmenter la robustesse en limitant la présence de systématiques,
- augmenter la puissance de détection en magnifiant les sources.

7.2.1 Soustraction robuste du continuum

La première étape cherche à supprimer les fortes sources de nuisance et certaines erreurs systématiques du cube. Pour cela on choisit d'estimer puis de soustraire les lignes de base des spectres explorés. En effet le signal recherché a pour principale signature une raie spectrale. La prise en compte du continuum (souvent bien plus intense chez les galaxies voisines que la raie recherchée) ne ferait en effet que gêner le processus de détection. La méthode classiquement utilisée au sein du consortium MUSE pour estimer ce continuum spectral est l'application d'un filtre médian (de fenêtre suffisamment grande par rapport à la taille caractéristique d'une raie). L'estimation au plus juste de ce continuum apparaissant comme cruciale afin d'éviter toute perte

du signal d'intérêt, des travaux ont été menés pour obtenir une méthode plus fine d'estimation de ce continuum. Ces travaux, publiés dans Bacher et al. 2016a et présentés en annexe B reposent sur une régression robuste locale permettant une estimation non-paramétrique de la ligne de base.

Dans le paragraphe 7.3, on comparera donc l'utilisation de cette méthode d'estimation avec l'utilisation du filtre médian lors de l'étape de soustraction du continuum.

7.2.2 Estimation robuste des paramètres de bruit

Les méthodes décrites dans le chapitre 6 nécessitent que le bruit soit symétrique et centré. Suite à la soustraction du continuum, on peut tout d'abord effectuer une première réduction à l'aide du cube de variance fourni avec les données, afin de rendre le bruit stationnaire (notons que cette stationnarité n'est pas nécessaire pour l'utilisation de la méthode COMET). On effectue ensuite un centrage robuste et une réduction plus fine, feuillet à feuillet à l'aide d'une approche de σ-clipping robuste développée dans [Meillier et al. 2017]. Après la réduction, on fait l'hypothèse que le bruit est stationnaire et réduit sur chaque feuillet spectral.

7.2.3 Filtrage adapté à la FSF

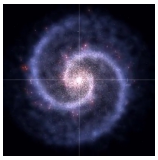
Du fait de l'influence de la FSF, le signal recherché est fortement dilué spatialement. La FSF étant principalement due aux turbulences atmosphériques, elle est indépendante du bruit de mesure de l'instrument. On peut donc faire l'hypothèse que le bruit dans le modèle d'observation (6.1) n'est pas filtré par la FSF. Une convolution spatiale des données par la FSF permet alors d'améliorer le RSB du halo. Notons qu'il ne s'agit pas à strictement parler d'un filtrage adapté au halo recherché puisque l'extension spatiale du halo est inconnu. Toutefois, le prix à payer est tout d'abord un élargissement artificiel de l'extension spatiale du halo. En pratique, le halo a une extension plus large que la FSF avec un profil d'intensité qui, tout comme celui de la FSF, décroît rapidement vers zéro sur les bords de son support. Cet élargissement artificiel peut ainsi être négligé.

Toutefois, un autre problème, comme nous l'avons vu dans le chapitre précédent, est que le régime asymptotique pour la méthode COMET est d'autant plus difficile à atteindre que les pixels sont fortement corrélés (la méthode BH empirique échappe à ce problème du fait de la possibilité d'apprentissage de la loi sur un grand nombre de données). Afin de pouvoir toujours assurer le contrôle avec COMET, tout en conservant un gain de RSB, le filtre appliqué par la suite correspond à la FSF tronquée à un noyau de taille 3 par 3. Cela permet de s'assurer que les corrélations induites sont de suffisamment faible portée par rapport aux cibles recherchées.

Il est à noter que les prochains jeux de données produits par MUSE profiteront de la mise en place en 2017 d'un système d'optique adaptative qui devrait permettre d'atténuer fortement la dégradation spatiale due à la FSF (au prix toutefois d'une FSF éventuellement non stationnaire dans le champ).

7.3 Résultats

Les résultats de détection sur les différents objets sélectionnés, pour les deux méthodes proposées ici et la méthode par champ de Markov triplet issue de [Courbot et al. 2016] sont montrés sur les figures 7.1 et 7.2.



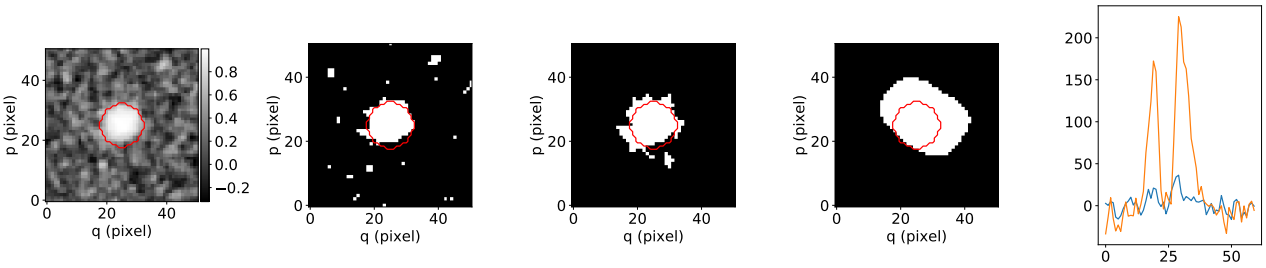
La première colonne présente les cartes de statistiques de max-test. La deuxième colonne montre les résultats produits par la procédure BH empirique pour un niveau nominal $q = 0.1$ de contrôle du FDR. La troisième colonne montre les résultats produits par l'approche COMET pour un niveau nominal $q = 0.1$ de contrôle du FDR. La quatrième colonne montre les résultats obtenus par la méthode par champ de Markov triplets proposée dans [Courbot et al. 2016]. La dernière colonne affiche les spectres estimés au sein du support de la FSF (spectre orange) et au sein du support détecté hors FSF (spectre bleu). Le cercle rouge indique la largeur de la FSF à la longueur d'onde de la raie considérée (le support est tronqué à 1% de la valeur centrale). Sur ces cartes, il apparaît clairement que plusieurs objets présentent des extensions asymétriques qui s'étendent au-delà la FSF (qui correspond, rappelons le, à l'étalement d'une source ponctuelle, i.e. la galaxie centrale ici). On peut voir que, comme sur simulation, la méthode COMET produit des cartes de détection *a priori* plus pertinentes (moins de pixels isolés correspondant vraisemblablement à des pics de bruit, détection plus étendue autour de la galaxie). Les résultats ne sont pas aisément comparables avec les résultats de classification par champ de Markov qui ont tendance soit à détecter des surfaces plus étendues et fortement régularisées, soit à ne rien détecter dans certains cas. Notons toutefois que certains objets, notamment le 547, présentent des extensions similaires dans les deux approches. Par ailleurs l'étude des spectres semble indiquer que les halos détectés possèdent bien la signature spectrale recherchée (bien que beaucoup plus faible que la galaxie, comme attendu). Cela permet de valider qualitativement le fait que les extensions observés sont bien dues à des émissions Lyman-α.

Ces résultats (ainsi que l'ensemble des codes associés) ont été transmis et sont désormais analysés au Centre de Recherche Astrophysique de Lyon.

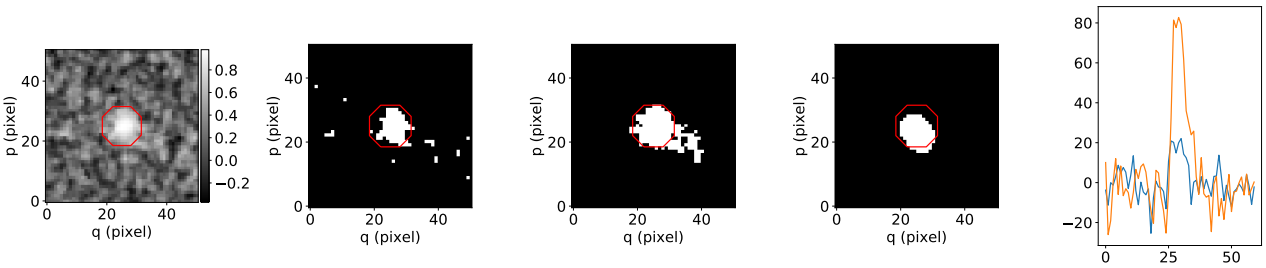
La figure 7.3 permet d'illustrer l'impact des méthodes de soustraction du continu étudiées. La méthode de régression robuste décrite en annexe B est comparée à une approche simple de filtre médian. On peut voir que les différences sont la plupart du temps minimes (à l'exception notable de l'objet 180 dont tout une extension disparaît lors de l'utilisation de l'estimation par régression robuste). Étant donné qu'il est très difficile de trancher entre les deux approches pour cette application en l'absence de vérité-terrain, la méthode retenue par défaut dans l'implémentation logicielle est le filtre médian pour des raisons de coût calculatoire.

Implémentation et coût calculatoire

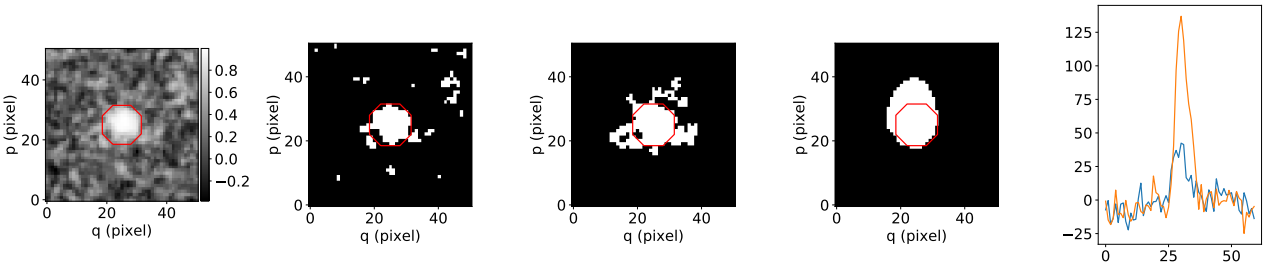
Le code a été développé en Python, en s'appuyant sur les bibliothèques *numpy* et *scipy*, ainsi que sur la bibliothèque *mpdaf* du consortium MUSE. Actuellement, le traitement d'une région 51×51 après prétraitement prend de l'ordre de 400ms pour les deux approches (procédure BH empirique et COMET) sur un processeur Intel à 8 coeurs cadencés à 3GHz. La majeure partie du coût calculatoire est due aux prétraitements qui peuvent être effectués en amont.



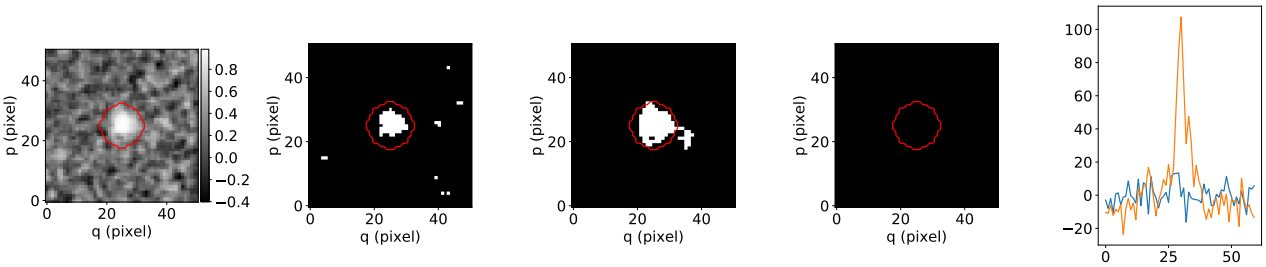
(a) Objet 106



(b) Objet 171

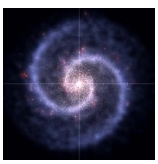


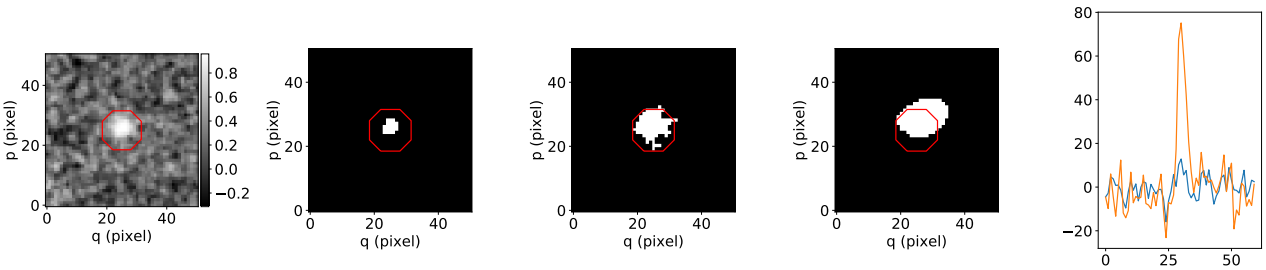
(c) Objet 180



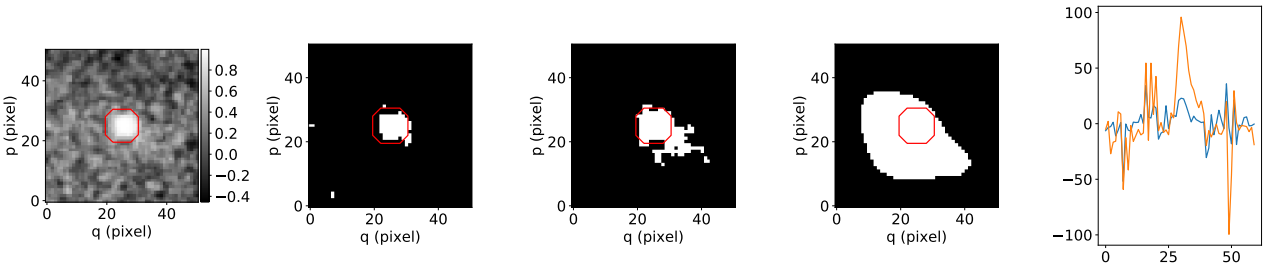
(d) Objet 218

FIGURE 7.1 – Résultats de détection sur plusieurs objets de l'UDF-10 (la numérotation correspond au catalogue défini dans [BACON et al. 2017]. Pour chaque ligne, de gauche à droite : carte des statistiques de test ; détection par l'algorithme 5 pour un niveau nominal $q = 0.1$ de contrôle du FDR ; détection par l'approche COMET pour un niveau nominal $q = 0.1$ de contrôle du FDR ; détection par champs de Markov triplets ; spectres estimés au sein du support de la FSF (spectre orange) et au sein du support détecté par COMET hors FSF (spectre bleu). Le cercle rouge indique la largeur de la FSF à la longueur d'onde de la raie considérée (le support est tronqué à 1% de la valeur centrale). Pour les deux méthodes proposées, la soustraction du continu est faite par filtre médian et une convolution par une FSF tronquée à un noyau 3 par 3 est appliquée en prétraitement.

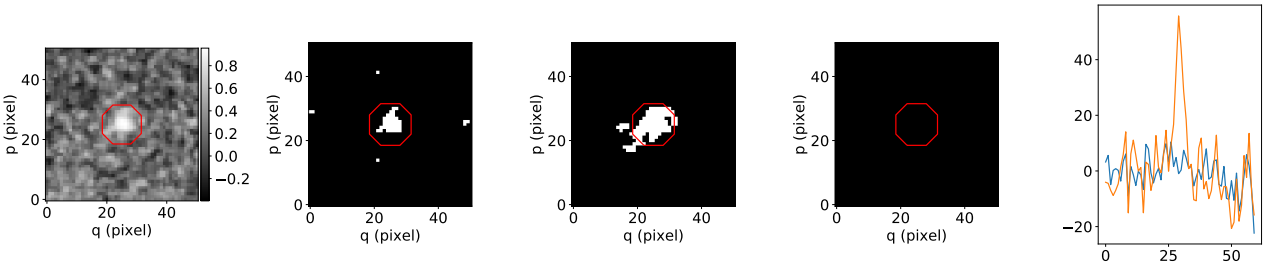




(a) Objet 364

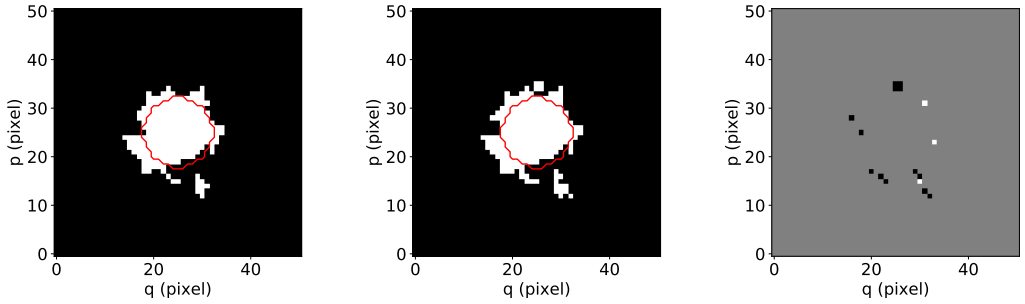


(b) Objet 547

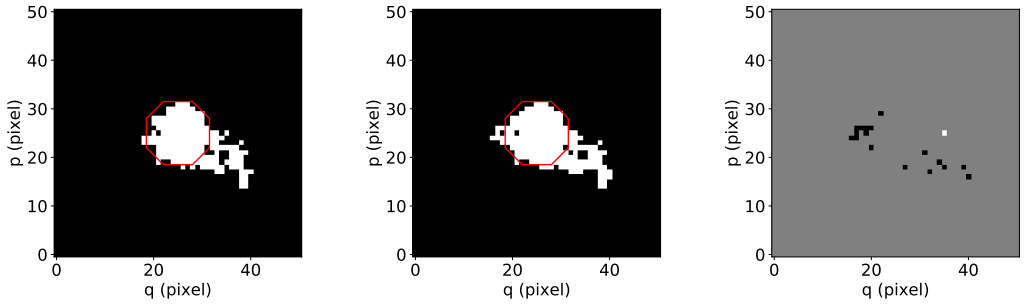


(c) Objet 6336

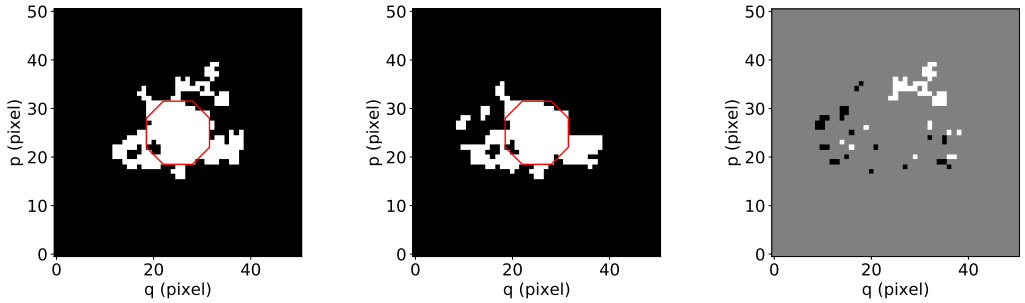
FIGURE 7.2 – Résultats de détection sur plusieurs objets de l'UDF-10 (suite) : voir légende de la figure 7.1



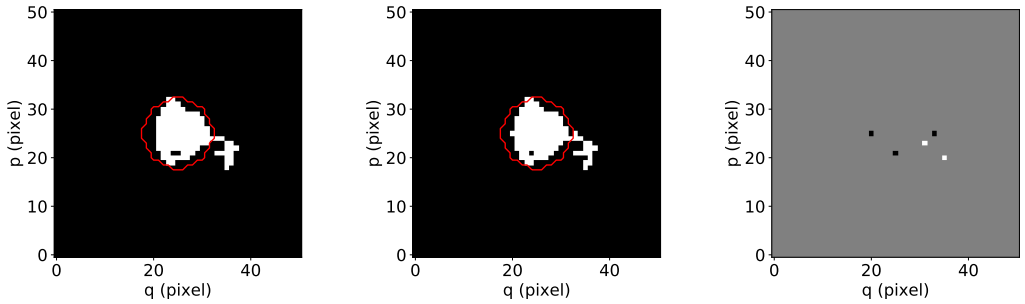
(a) Objet 106



(b) Objet 171

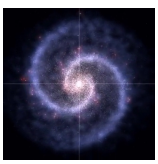


(c) Objet 180



(d) Objet 218

FIGURE 7.3 – Comparaisons des méthodes de soustraction de continu. A gauche : le résultat de détection en utilisant le filtre médian comme estimateur du continuum ; au centre : le résultat de détection en utilisant la méthode de régression robuste développée en annexe B ; à droite : la différence entre les deux (les pixels noirs indiquent les détections obtenues uniquement en utilisant la méthode de régression robuste, les pixels blancs indiquent les détections obtenues uniquement en utilisant le filtre médian.



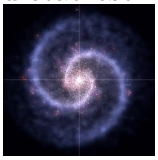
Bilan et perspectives de la partie II

Dans cette partie, nous avons étudié le problème de la détection d'une source spatialement étendue, caractérisée par une signature spectrale faible et potentiellement variable, au sein d'un fond difficile à modéliser. Pour résoudre ce problème, un détecteur non-supervisé est proposé, fondé sur une approche de type max-test, comme développé dans [Arias-Castro et al. 2011]. Ce détecteur permet de prendre en compte une possible variabilité de la source à l'aide d'un dictionnaire redondant. Il ne nécessite pas de modélisation paramétrique du fond mais s'appuie principalement sur une hypothèse de symétrie du bruit couplée à une décomposition parcimonieuse non-négative du signal d'intérêt. Cela permet d'estimer la distribution de la statistique de test et d'implémenter une procédure de détection simple mais robuste aux erreurs de modélisations. On peut ensuite appliquer une procédure BH empirique de contrôle du FDR à l'aide de cette distribution empirique des statistiques de max-test, afin de garantir un contrôle global des erreurs. Pour améliorer la puissance de détection, tout en assurant le même contrôle du FDR, une procédure alternative de contrôle a été développée, COMET, s'inspirant de la récente procédure BC de contrôle du FDR. Cette procédure permet d'exploiter ici un *a priori* de connexité spatiale de la cible recherchée, en s'appuyant (comme la méthode d'apprentissage de la distribution du bruit) sur la symétrie du bruit et la positivité de la cible (tout en relâchant les contraintes de stationnarité du bruit).

Les deux approches sont testées sur les données MUSE de l'UDF-10 et comparées aux résultats produits par la méthode de classification par champ de Markov développée également pour l'application MUSE dans [Courbot et al. 2016]. Du point de vue méthodologique, les résultats sur données simulées permettent de montrer la supériorité de l'approche COMET sur la procédure BH empirique. Du point de vue applicatif, un certain nombre des objets étudiés semblent posséder des extensions spatiales bien plus importantes que l'étendue de la FSF correspondant à l'étendue d'une simple source ponctuelle. Ces résultats ont été transmis au consortium MUSE et une analyse astrophysique de ces extensions spatiales détectées est en cours. La comparaison avec l'approche par champ de Markov semble montrer des détections parfois beaucoup plus étendues et très régularisées pour les champs de Markov. Toutefois cette approche se traduit parfois par une absence complète de détection, là où COMET détecte toujours au moins le cœur de la galaxie. Rappelons également qu'un point clé des méthodes proposées ici est de permettre un contrôle précis du taux de fausses découvertes et donc de régler en toute connaissance de cause la sensibilité du test de détection.

Perspectives

L'approche COMET est très générique et de nombreuses implémentations sont envisageables, que ce soit en changeant les statistiques de test ou en changeant la procédure de sélection, afin de prendre en compte d'autres informations *a priori*. Dans le cadre de l'application MUSE, une extension particulièrement intéressante consiste à prendre en compte le fait que la signature spectrale varie lentement d'un pixel à l'autre. On peut alors très facilement introduire dans la procédure de sélection une nouvelle contrainte de continuité sur la forme de la signature à détecter, à l'aide par exemple d'un modèle d'état et d'un suivi par filtrage Kalman. Ceci permettrait un gain significatif en puissance dès lors que le modèle d'état est



adapté. La principale difficulté sera de conserver une procédure de sélection sans biais afin de conserver le contrôle des erreurs.

D'autres implémentations permettraient également de s'appliquer à d'autres contextes que les données MUSE. On peut par exemple envisager de remplacer la procédure de sélection promouvant la connexité spatiale par une procédure promouvant une continuité temporelle dans des données de séries temporelles.

Notons également que la mise en place en 2017 d'un système d'optique adaptative sur le télescope de MUSE va radicalement changer la qualité spatiale des données produites. Cela permettra ainsi d'améliorer fortement le RSB, supprimant ainsi l'utilisation peu satisfaisante, dans un contexte de détection, du filtrage adapté par la FSF. Il sera donc très intéressant d'appliquer à nouveau les approches développées ici sur ces nouvelles données.

Conclusion et perspectives

L'objectif de cette thèse est de proposer des solutions méthodologiques à deux problématiques apparaissant dans l'analyse des champs profonds hyperspectraux issus de l'instrument MUSE. La première problématique consistait à estimer le spectre des sources présentes dans le champ malgré les phénomènes de mélange spectral dû à un manque de résolution spatiale dans les données MUSE observées. Cela permet en effet ensuite aux astrophysiciens d'effectuer des analyses des propriétés physico-chimiques des différentes sources galactiques. La seconde problématique est de pouvoir détecter des sources faibles spatialement étendues. On estime en effet que la plupart des jeunes galaxies sont entourées de halos de gaz circum-galactiques. Ces halos de gaz émettent une signature spectrale concentrée sur une raie de quelques feuillets spectraux.

Démélange spectral

La problématique de démélange spectral des sources MUSE a été abordée en combinant les données MUSE avec les données du télescope spatial Hubble, de bien meilleure résolution spatiale (mais de résolution spectrale bien plus faible). Les données MUSE sont modélisées comme la dégradation spatiale d'un mélange linéaire de sources, et les données HST sont vues comme la dégradation spectrale de ce même mélange. L'information des données HST a été transférée à la résolution MUSE afin d'écrire un système linéaire à inverser. Cette inversion doit toutefois être régularisée car le problème se trouve rapidement mal conditionné lorsque les sources sont proches spatialement. La régularisation a été choisie de façon à prendre en compte les spécificités des données étudiées : les spectres pouvant se décomposer en une ligne de base et un ensemble de raies parcimonieux, les données ont été décomposées selon ce principe et chaque composante a été estimée avec une régularisation adaptée. Cette méthode a alors été testée sur les données MUSE de l'UDF-10 avec des résultats probants.

Détection de sources étendues

Pour répondre à la problématique de détection de sources étendues, nous avons mis en place une méthode de détection s'appuyant sur la construction de statistiques de max-test sur un dictionnaire redondant. Nous avons ensuite proposé une méthode d'estimation de la distribution de la statistique de test et de la proportion π_0 de données sous $\mathcal{H}_0$, à l'aide des propriétés de symétrie du bruit couplée à la positivité de la contribution du signal recherché. On peut ensuite appliquer une procédure BH empirique de contrôle du FDR à l'aide de cette distribution empirique des statistiques de max-test, afin de garantir un contrôle global des erreurs. La méthode proposée ne nécessite ainsi aucune autre spécification sur le modèle de bruit que la symétrie. Pour améliorer la puissance de détection, tout en assurant le même contrôle du FDR, une procédure alternative de contrôle a été développée, COMET, s'inspirant de la récente procédure BC de contrôle du FDR. Cette procédure permet de généraliser l'approche précédente en exploitant ici un *a priori* de connexité spatiale de la cible recherchée, sous les mêmes hypothèses de symétrie du bruit et de positivité de la cible. Des conditions théoriques de

contrôle exact et asymptotique du FDR ont été données. Les deux approches ont été testées sur les données MUSE de l'UDF-10 avec des résultats particulièrement probants pour la méthode COMET.

Principales contributions

Concernant la problématique de démélange, les principales contributions sont les suivantes :

- Implémentation d'une chaîne de conversion des données HST vers MUSE.
- Mise en place d'un ensemble de régularisations adapté à la nature des données MUSE (spectres composés d'un continuum et de raies parcimonieuses)

Concernant la problématique de détection de sources étendues, les principales contributions sont les suivantes :

- Définition d'une méthode simple d'apprentissage robuste de la distribution des statistiques de test pour le problème considéré (bruit symétrique et sources à contribution positive).
- Conception d'un algorithme générique de contrôle du FDR avec prise en compte de connexité spatiale, à l'aide des mêmes hypothèses de symétrie du bruit symétrique et positivité des sources.
- Construction d'un estimateur de ligne de base spectrale par une régression locale robuste et non paramétrique (voir annexe B).

Notons que ces développements méthodologiques sont allés de pair avec une prise en main avancée des données réelles, assez chronophage, qui a fortement influencé le choix et le réglage de ces méthodes pour s'assurer qu'elles soient les plus efficaces possibles. Les méthodes proposées ont notamment été systématiquement testées sur un grand nombre de cas issues des données réelles, afin de s'assurer également de la robustesse de leur implémentation. Soulignons également l'intérêt du choix du FDR comme outil de contrôle des erreurs. Cet outil nous semble en effet particulièrement adapté aux problèmes de détections de sources multi-pixeliques comme abordés ici, en fournissant une notion très concrète du taux de fausses découvertes.

Production logicielle

Les méthodes développées au cours de cette thèse ont été implémentées de façon à s'intégrer dans la bibliothèque logicielle développée au sein du consortium MUSE (Bacon et al. 2016). Les codes ont été transmis au consortium et visent désormais à être utilisé de façon routinière, et en toute autonomie, par le consortium et la communauté astrophysique.

Communication scientifique

Comme décrit en introduction, ces travaux ont été diffusés sous la forme de plusieurs conférences internationales et un article de revue, principalement à destination de la communauté du traitement du signal. Des communications visant plus particulièrement la communauté astrophysique sont en préparation.

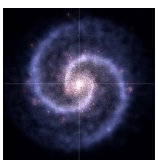
Perspectives

Les perspectives à ces travaux sont nombreuses, pour les deux grandes problématiques abordées. Etant déjà exposées en détail dans les bilans de chaque partie, elles ne sont que rappelées brièvement ici.

Concernant le démélange spectral des sources, le pouvoir de séparation de la méthode proposé est limitée par l'information spatiale fournie par HST. Une approche exploitant conjointement l'information spectrale et spatiale pour le démélange (par exemple via une méthode de super-résolution par fusion de données) permettrait de résoudre certains cas. Cela pourrait être facilité par la prise en compte d'informations physiques sur les spectres, à l'aide de modèles de spectres. On peut d'ailleurs penser que l'analyse des spectres dans les données MUSE va permettre d'enrichir de façon conséquente les modèles et dictionnaires existants de spectres galactiques.

Du point de vue de la détection de sources étendues, plusieurs extensions sont également envisageables. La méthode COMET est très générique et peut notamment être adaptée pour prendre en compte de nouvelles informations *a priori*, comme l'évolution de la signature spectrale de la cible. Par ailleurs, des travaux exploratoires pour aborder ce problème d'un point de vue de l'apprentissage automatique (*machine learning*) ont également été menés et sont présentés dans l'annexe C. Ces travaux sont pour l'instant encore très préliminaires et n'offrent pour l'instant pas les mêmes garanties de contrôle que les méthodes proposées dans la partie II. De nombreuses améliorations sont donc certainement possibles ainsi qu'une comparaison systématique avec les résultats obtenues par COMET, qui n'a pas encore été menée.

Notons également que la mise en place d'un système d'optique adaptative sur le télescope de MUSE (en cours de test) va fortement impacter les deux problématiques. Concernant le démélange, cela devrait permettre de résoudre immédiatement un certain nombre de situations de mélanges. Toutefois, la nouvelle FSF sera vraisemblablement plus difficile à modéliser, ce qui nécessitera donc de nouveaux développements adaptés afin de résoudre les situations de mélange restantes. Concernant la problématique de détection, l'optique adaptative se traduit par un gain important en terme de RSB, ce qui devrait fortement profiter aux méthodes de détection proposées.



Annexe A

Preuves

Preuve du lemme 1 (page 90)

Lemme (Estimateur empirique de la médiane sous l'hypothèse nulle)
Notons $t_{(1)} < t_{(2)} < \ldots < t_{(2n)}$ *les valeurs ordonnées des statistiques de l'échantillon* $\boldsymbol{t} = (T_{max}(\boldsymbol{y}_1), \ldots, T_{max}(\boldsymbol{y}_n), -T_{min}(\boldsymbol{y}_1), \ldots, -T_{min}(\boldsymbol{y}_n))$. *Notons* $\widehat{\mu}_0$ *la médiane empirique de* $\boldsymbol{t}$*, définie par*

$$\widehat{\mu}_0 = \frac{t_{(n)} + t_{(n+1)}}{2}. \tag{6.9}$$

Alors, lorsque les hypothèses A3 et A4 sont valides, $\widehat{\mu}_0$ *vérifie l'équation* (6.8) *et est un estimateur consistent de la médiane sous l'hypothèse nulle* μ_0*.*

Démonstration. Notons $\boldsymbol{g} = (T_{\max}(\boldsymbol{y}_1), \ldots, T_{\max}(\boldsymbol{y}_n))$ l'ensemble des n statistiques des maxima, dont les éléments sont notés g_i pour $1 \leq i \leq n$. De manière semblable, $\boldsymbol{s} = (-T_{\min}(\boldsymbol{y}_1), \ldots, -T_{\min}(\boldsymbol{y}_n))$ est l'ensemble des n statistiques des opposés des minima, dont les éléments sont notés s_i pour $1 \leq i \leq n$.

On montre d'abord que $\widehat{\mu}_0$ vérifie l'équation (6.8), c'est à dire que

$$\#\{g_i \leq \widehat{\mu}_0\} = \#\{s_i > \widehat{\mu}_0\}.$$

Pour des distributions continues, $\Pr(t_{(n)} = t_{(n+1)}) = 0$. Donc, d'après (6.9), on obtient que

$$\#\{t_i \leq \widehat{\mu}_0\} = n$$

avec une probabilité de 1. L'ensemble $\boldsymbol{t}$ est l'union de $\boldsymbol{g}$ et $\boldsymbol{s}$: si $m_0 = \#\{g_i \leq \widehat{\mu}_0\}$, alors $\#\{s_i \leq \widehat{\mu}_0\} = n - m_0$. En conséquence,

$$\#\{s_i > \widehat{\mu}_0\} = n - (n - m_0) = m_0 = \#\{g_i \leq \widehat{\mu}_0\},$$

ce qui montre que $\widehat{\mu}_0$ vérifie l'équation (6.8).

Montrons à présent que $\widehat{\mu}_0$ converge en probabilité vers μ_0. Puisque $\widehat{\mu}_0$ vérifie (6.8), et que $\overline{F}(t)$ (resp. $\overline{G}(t)$) converge en probabilité vers $F(t)$ (resp. $G(t)$) pour tout $t \in \mathbb{R}$, alors $\widehat{\mu}_0$ converges en probabilité vers la solution de

$$F(t) = G(t).$$

si cette équation admet une solution unique. En supposant que la médiane de F_0 est définie de façon unique, on en déduit que pour $t > \mu_0$

$$F(t) = \pi_0 F_0(t) + \pi_1 F_1(t) \geq \pi_0 F_0(t) > \pi_0 F_0(\mu_0) = \pi_0/2,$$

De plus, pour $t > \mu_0$

$$G(t) = \pi_0 G_0(t) < \pi_0 G_0(\mu_0) = \pi_0/2 < F(t),$$

où la première inégalité est due à l'hypothèse "bruit seul" A4. En conséquence, il n'existe pas de solution sur $(\mu_0, +\infty)$.
De façon semblable, selon l'hypothèse "bruit seul" A3, pour $t < \mu_0$, on a $F(t) < G(t)$. Donc la solution unique est $t = \mu_0$, où

$$F(\mu_0) = \pi_0 F_0(\mu_0) = \pi_0/2 = \pi_0 G_0(\mu_0) = G(\mu_0),$$

ce qui conclut la preuve. □

Preuve de la proposition 2 (page 90)

Proposition (Estimateurs empiriques sous $\mathcal{H}_0$)
Sous les hypothèses A3 et A4 (hypothèses de bruit seul),

$$\widehat{\pi}_0 = \min\left\{2n_0/n, 1\right\}, \tag{6.10}$$

est un estimateur consistant de la proportion π_0 d'échantillons sous $\mathcal{H}_0$, et

$$\widehat{F}_0(t) = \frac{\#\left\{s_{0,i} \leq t\right\} + \#\left\{g_{0,i} \leq t\right\}}{2n_0} \tag{6.11}$$

est un estimateur consistant de la distribution sous l'hypothèse nulle $F_0(t)$, pour $t \in \mathbb{R}$.

Démonstration. On montre d'abord que l'estimateur de π_0 donné dans l'équation (6.10) est consistant. D'après le lemme 1, $\widehat{\mu}_0$ converge en probabilité vers μ_0 : $\widehat{\mu}_0 \xrightarrow{P} \mu_0$. L'inégalité triangulaire assure que

$$|\overline{F}(\widehat{\mu}_0) - F(\mu_0)| \leq |\overline{F}(\widehat{\mu}_0) - F(\widehat{\mu}_0)| + |F(\widehat{\mu}_0) - F(\mu_0)|.$$

Le premier terme de droite est dominé par $\sup_t |\overline{F}(t) - F(t)|$, qui converge en probabilité vers 0, selon l'hypothèse A5. Le second terme converge aussi en probabilité vers 0, d'après le théorème de Mann-Wald (*continuous mapping theorem*, Vaart 1998, p7). D'où

$$\overline{F}(\widehat{\mu}_0) \xrightarrow{P} F(\mu_0).$$

D'après l'équation (6.7),

$$F(\mu_0) = \pi_0 F_0(\mu_0) = \frac{\pi_0}{2}.$$

Comme $2\overline{F}\left(\widehat{\mu}_0\right) = 2\frac{n_0}{n}$, on a que

$$\widetilde{\pi}_0 \equiv 2\frac{n_0}{n} \xrightarrow{P} \pi_0 \in]0, 1].$$

De plus $\widehat{\pi}_0 = \min\left\{\widetilde{\pi}_0, 1\right\}$ converge également en probabilité vers π_0.

On montre désormais la consistance de l'estimateur (6.11) pour $t \in \mathbb{R}$. En utilisant (6.7) et l'hypothèse A5, on obtient que $\overline{F}(t) \xrightarrow{P} \pi_0 F_0(t)$ pour tout $t \leq \mu_0$. Alors, d'après le théorème de Slutsky [1],

$$\overline{F}(t)/\widetilde{\pi}_0 \xrightarrow{P} F_0(t).$$

Puisque $\widehat{F}_0(t) = \overline{F}(t)/\widetilde{\pi}_0$ pour $t \leq \mu_0$, l'estimateur est consistant pour $t \leq \mu_0$. La démonstration pour $t > \mu_0$ peut se faire de manière similaire, en notant que $\widehat{F}_0(t) = 1 - \overline{G}(t)/\widetilde{\pi}_0$ pour $t > \mu_0$. □

Preuve de la proposition 3 (page 94)

Proposition (Estimateur de Storey de π_0)
L'estimateur empirique $\widehat{\pi}_0$ défini dans l'équation (6.10) et l'estimateur de Storey $\hat{\pi}_0^*(\zeta)$ dérivé des p-valeurs empiriques définies dans (6.12) sont égaux pour tout $\zeta = \frac{k}{2n_0}$ avec $k \in \{n_0, \ldots, 2n_0 - 1\}$, et sont asymptotiquement équivalent pour tout $\zeta \in [\frac{1}{2}, 1[$.

Démonstration. Notons $T_{\max}(\boldsymbol{y}_{(1)}) \leq T_{\max}(\boldsymbol{y}_{(2)}) \leq \cdots \leq T_{\max}(\boldsymbol{y}_{(n)})$ les statistiques de max-test *ordonnées*, et $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}$ les p-valeurs *ordonnées* (alors que p_i indique la p-valeur associée au pixel i). D'après (6.12), on obtient que

$$p_{(i)} = 1 - \widehat{F}_0\left(T_{\max}(\boldsymbol{y}_{(j)})\right) \quad \text{for } 1 \leq i \leq n,$$

où $j = n - i + 1$.
Pour $j \leq n_0$, on a $T_{\max}(\boldsymbol{y}_{(j)}) \leq \mu_0$ donc

$$\#\left\{s_{0,i} \leq T_{\max}(\boldsymbol{y}_{(j)})\right\} = j$$

et

$$\#\left\{g_{0,i} \leq T_{\max}(\boldsymbol{y}_{(j)})\right\} = 0.$$

D'où, $2n_0\widehat{F}_0\left(T_{\max}(\boldsymbol{y}_{(j)})\right) = j$ pour $j \leq n_0$, soit $2n_0 p_{(i)} = 2n_0 - n + i - 1$ pour $i \geq n - n_0 + 1$. Pour $k \geq n_0$, notons $i_k = n - 2n_0 + 1 + k$. Alors $i_k \geq n - n_0 + 1$ donc $2n_0 p_{(i_k)} = 2n_0 - n + i_k - 1 = k$. D'où

$$\#\{2n_0 p_i > k\} = n - i_k = 2n_0 - k - 1.$$

Si $\zeta = k/2n_0$, alors

$$\#\{p_i > \zeta\} = \#\{2n_0 p_i > k\}.$$

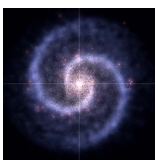
De plus

$$\frac{1 + \#\{p_i > \zeta\}}{(1-\zeta)n} = \frac{2n_0 - k}{(1 - k/2n_0)n} = \frac{2n_0}{n},$$

ce qui montre que $\hat{\pi}_0^*(\zeta) = \widehat{\pi}_0$.

Dans le cas général où $\zeta \in [\frac{1}{2}, 1[$, ζ et $k_\zeta/2n_0$, où $k_\zeta = \lfloor 2n_0\zeta \rfloor \in \{n_0, \ldots, 2n_0 - 1\}$, sont asymptotiquement équivalents lorsque n_0 tend vers l'infini. De plus, $\hat{\pi}_0^*(\zeta)$ et $\hat{\pi}_0^*(k_\zeta/2n_0) = \widehat{\pi}_0$ sont asymptotiquement équivalents. Cela conclut la preuve. □

1. Soient $\{X_n\}$, $\{Y_n\}$ des suites de variables aléatoires à valeur respectivement dans $\mathbb{R}^p$ et $\mathbb{R}^q$. Si X_n converge en loi vers X, et si Y_n converge en probabilité vers une constante c, alors le couple (X_n, Y_n) converge en loi vers le couple (X, c).



Preuve de la proposition 7 (page 114)

Proposition
Pour tout $t \in \mathbb{R}$ et $m \geq 2$, on définit $M_{m+1}(t)$ de façon récursive sous $\mathcal{H}_0$ par

$$M_{m+1}(t) = \Pr\left(z_1^{m+1} \leq t \mid z_2^{m+1} \leq t, z_3^{m+1} \leq t\right) \times M_m(t), \tag{6.20}$$

avec $M_2(t) = \Pr\left(z_1^2 \leq t, z_2^2 \leq t\right)$. Sous les hypothèses mentionnées précédemment, une borne supérieure de la PFA α_m est donnée par $1 - M_m(\eta)$.

Démonstration. Sous $\mathcal{H}_0$, pour un seuil t on a :

$$\begin{aligned}\Pr\left(\max \boldsymbol{z}^{m+1} \leq t\right) &= \Pr\left(z_1^{m+1} \leq t, ..., z_{m+1}^{m+1} \leq t\right) \\ &= \Pr\left(z_1^{m+1} \leq t \mid z_2^{m+1} \leq t, ..., z_{m+1}^{m+1} \leq t\right) \\ &\times \Pr\left(z_2^{m+1} \leq t, ..., z_{m+1}^{m+1} \leq t\right)\end{aligned}$$

Comme $\boldsymbol{D}^m \geq 0$ et $\boldsymbol{y} \sim \mathcal{N}(0, \boldsymbol{I}_m)$ sous $\mathcal{H}_0$, $\boldsymbol{z}^m$ est positivement associé au sens de [Esary et al. 1967]. Donc,

$$\Pr\left(z_1^{m+1} \leq t \mid z_2^{m+1} \leq t, ..., z_{m+1}^{m+1} \leq t\right) \geq \Pr\left(z_1^{m+1} \leq t \mid z_2^{m+1} \leq t, z_3^{m+1} \leq t\right) \tag{A.1}$$

En utilisant les expressions numériques données dans [Genz et Bretz 2009], on peut calculer de façon précise le terme de droite de l'équation (A.1). Notons que ce terme donne une borne inférieure assez précise, car z_2^{m+1} et z_3^{m+1} sont les variables les plus corrélées à z_1^{m+1} parmi z_j^{m+1} pour $j \geq 2$.

De plus, par construction, les translations entre les atomes de $\boldsymbol{D}^{m+1}$ sont inférieures à celles entre les atomes de $\boldsymbol{D}^m$. Puisque la fonction d'autocorrélation est supposée être non-croissante en fonction des translations absolues, on obtient que le vecteur gaussien $\left(z_2^{m+1}, ..., z_{m+1}^{m+1}\right)$, de taille m, a des correlations plus grandes que le vecteur gaussien, $(z_1^m, ..., z_m^m)$, toujours de taille m. Par hypothèse, ces deux vecteurs sont centrés avec une variance marginale unitaire sous $\mathcal{H}_0$. De plus le lemme de Slepian (Slepian 1962) montre que :

$$\Pr\left(z_2^{m+1} \leq t, ..., z_{m+1}^{m+1} \leq t\right) \geq \Pr\left(z_1^m \leq t, ..., z_m^m \leq t\right). \tag{A.2}$$

En combinant (A.1) et (A.2), on peut alors minimiser $\Pr\left(\max \boldsymbol{z}^m \leq t\right)$ par une fonction $M_m(t)$ définie récursivement par :

$$M_{m+1}(t) = \Pr\left(z_1^{m+1} \leq t \mid z_2^{m+1} \leq t, z_3^{m+1} \leq t\right) \times M_m(t),$$

où $M_2(t) = \Pr\left(z_1^2 \leq t, z_2^2 \leq t\right)$. Cela donne une borne supérieure pour la PFA de la proposition 7. Notons que l'application numérique souligne que $M_m(t)$ augmente avec t. Il est alors possible d'inverser (numériquement) $M_m(\eta)$ afin d'obtenir η_m pour un niveau de contrôle α : $\eta_m = M_m^{-1}(1-\alpha)$ vérifie $\Pr(\max \boldsymbol{z}^m > \eta_m) \leq \alpha$. □

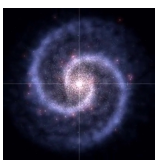
Annexe B

Méthode d'estimation du continuum spectral

Cette annexe présente une méthode de régression robuste développée pour permettre d'estimer de façon précise les lignes de base des spectres dans les données MUSE. Ces lignes de base correspondent aux émissions continues des galaxies et sont des sources de nuisance pour la détection des faibles sources étendues extragalactiques comme les halos. Comme détaillé dans l'article de synthèse [Komsta 2011], un grand nombre de méthodes ont été développées dans les dernières années pour résoudre le problème de l'estimation de ligne de base en spectroscopie. On peut citer notamment AIRPLS (Zhang et al. 2010), LOWESS (Cleveland 1981), ou plus récemment BEADS (Ning et al. 2014). AIRPLS utilise une approche non-paramétrique par moindres carrés pénalisés itératif (IRLS). BEADS s'appuie sur une approche par filtrage où la ligne de base est assimilée à la composante basse fréquence du signal. Toutefois la plupart de ces techniques ont été développées dans un contexte de chromatographie et ne sont pas adaptés aux signaux très bruités des données MUSE. De plus le grand nombre de spectres à traiter implique de choisir une méthode adaptative et non-supervisée

Il est par conséquent pertinent de chercher à mettre en place une nouvelle méthode d'estimation de la ligne de base adaptée aux spécificités des données MUSE. L'idée ici est de considérer les raies d'un spectre comme des valeurs aberrantes, dont il faut se prémunir pour estimer la ligne de base du spectre. On s'appuie alors sur la méthode de régression robuste *least trimmed squares* ou LTS (Rousseeuw et Leroy 1987, p.15) qui consiste à ne pas prendre en compte dans la régression les points ayant les plus grands résidus. La méthode LTS nécessite toutefois de définir la proportion de valeurs aberrantes à rejeter. Afin de faire face à ce problème, une approche adaptative simple est proposée, basée sur la distribution empirique des résidus ordonnés. On obtient donc une nouvelle méthode nommée *adaptive least trimmed squares* (ALTS) qui permet de conserver la robustesse de l'approche LTS classique tout en s'adaptant à la proportion de valeurs aberrantes, améliorant ainsi son efficacité asymptotique en l'absence de valeurs aberrantes. Il est à noter qu'une autre approche LTS adaptative a été proposée dans [Xu et al. 2014] mais elle s'appuie sur un critère adapté à un contexte bien précis de comparaisons par paires, qui n'est donc pas applicable à notre objectif final de détection d'une ligne de base.

A partir de cette approche LTS adaptative, la méthode d'estimation de ligne de base est construite en suivant l'idée générale de l'algorithme LOWESS algorithm, une méthode classique de lissage. Il s'agit de réaliser une régression linéaire locale par fenêtre glissante, la régression locale étant faite par l'approche LTS adaptative. Les résultats présentés ici s'appuient sur les travaux publiés dans [Bacher et al. 2016a].



B.1 L'algorithme ALTS

B.1.1 Estimation par LTS

On s'intéresse ici au problème de régression linéaire :

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{B.1}$$

où $\boldsymbol{y} \in \mathbb{R}^n$ est le vecteur d'observations, $\boldsymbol{X} \in \mathbb{R}^{n\times q}$ est une matrice de régression donnée, $\boldsymbol{\beta} \in \mathbb{R}^q$ est le vecteur des coefficients de régression et $\boldsymbol{\epsilon} \in \mathbb{R}^n$ est un bruit blanc centré de variance σ^2. L'estimateur classique des moindres carrés (LS) de $\boldsymbol{\beta}$ cherche à résoudre :

$$\widehat{\boldsymbol{\beta}}_{\text{LS}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 = \sum_{i=1}^{n} [r_i(\boldsymbol{\beta})]^2, \tag{B.2}$$

où $\boldsymbol{x}_i$ est le $i$ème vecteur colonne de $\boldsymbol{X}$, et $r_i(\boldsymbol{\beta}) \equiv y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}$ est $i$ème résidu, pour $i = 1, \ldots, n$.

Dans un cadre de régression robuste, on suppose à présent qu'une proportion des observations est composée de données atypiques ou aberrantes. Alors, pour $i = 1, \ldots, n$, le modèle d'observation devient $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i$ pour des données "normales" alors qu'il n'est pas défini pour les valeurs aberrantes. Par nature, les données aberrantes sont supposées être distribuées de façon très distinctes du modèle de prédiction $\boldsymbol{x}_i^T\boldsymbol{\beta}$. Cela génère des résidus importants qui peuvent dégrader l'estimateur au sens des moindres des carrés. L'estimateur LTS est un estimateur robuste classique pour faire face à ce problème en négligeant les résidus les trop importants.

Pour un $\boldsymbol{\beta}$ donné, notons $r_{(i)}$, pour $i = 1, \ldots, n$, les valeurs des résidus ordonnées de façon absolue, telles que $|r_{(1)}| \leq |r_{(2)}| \leq \ldots \leq |r_{(n)}|$. L'estimateur α-LTS est alors défini par (Rousseeuw et Leroy 1987) :

$$\widehat{\boldsymbol{\beta}}_{\text{LTS}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{h} [r_{(i)}(\boldsymbol{\beta})]^2, \tag{B.3}$$

où $\alpha \in [1/2, 1]$ et $h = [\alpha n]$ sont respectivement la proportion et le nombre des plus petits résidus retenus dans le critère à minimiser.

La définition donnée en (B.3) implique de résoudre un problème d'optimisation non-convexe pour calculer l'estimateur LTS. Toutefois, comme exposé dans [Xu et al. 2014], ce problème peut en fait être vu comme le problème d'optimisation sous contrainte suivant :

$$\begin{cases} \min_{\boldsymbol{\beta},\boldsymbol{w}} \sum_{i=1}^{n} w_i (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2, \\ \text{avec } ||\boldsymbol{w}||_0 \geq h,\ \boldsymbol{w} \in \{0,1\}^n \end{cases} \tag{B.4}$$

L'optimisation se fait alors conjointement sur $\boldsymbol{\beta}$ et sur le vecteur de poids binaires $\boldsymbol{w} \in \{0,1\}^n$. Cela peut être résolu par une approche itérative alternant deux étapes simples. Pour un $\boldsymbol{w}$ fixé, le minimiseur global sur $\boldsymbol{\beta}$ est donné par :

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}, \tag{B.5}$$

où $\boldsymbol{W} = \text{diag}\,(\boldsymbol{w})$ est une matrice de pondération diagonale. Pour un $\boldsymbol{\beta}$ fixé, le minimiseur global sur $\boldsymbol{w}$ est le vecteur binaire tel que, pour $i = 1, \ldots, n$:

$$w_i = \begin{cases} 1 & \text{si } |r_i(\boldsymbol{\beta})| \leq |r_{(h)}(\boldsymbol{\beta})|, \\ 0 & \text{sinon} \end{cases} \tag{B.6}$$

Algorithme 7 LTS

1: *Entrée :* $\boldsymbol{X}$ (matrice de régresseurs), $\boldsymbol{y}$ (données), h (nombre de données non aberrantes), $\widehat{\boldsymbol{\beta}}^0$ (initialisation)

2: *Boucle :*
3: **while** $\widehat{\boldsymbol{\beta}}^k \neq \widehat{\boldsymbol{\beta}}^{k-1}$ **do**
4: Calcul de $\widehat{\boldsymbol{w}}^k$ à l'aide de $\widehat{\boldsymbol{\beta}}_{k-1}$ et (B.6) ▷ Mise à jour
5: Calcul de $\widehat{\boldsymbol{\beta}}^k$ à l'aide de $\widehat{\boldsymbol{w}}^k$ et (B.5) ▷ Prédiction

6: *Sortie :* $\widehat{\boldsymbol{\beta}}^k$

On obtient finalement l'algorithme itératif décrit en Alg. 7 : Cela correspond à une méthode classique de prédiction-correction, où chaque étape fait décroître le critère donné en (B.4). Le vecteur binaire $\boldsymbol{w}$ appartenant à un espace avec un nombre fini d'états, la convergence vers un minimum local du problème initial (B.3) est donc garantie après un nombre fini d'itérations. Comme décrit dans [Rousseeuw et Van Driessen 2006], il est possible, comme dans l'algorithme k-means, de choisir aléatoirement plusieurs sous-ensemble d'observations, et ainsi réaliser plusieurs initialisations aléatoires de l'algorithme. Cela peut permettre d'éviter d'être piégé dans un minimum local, par exemple lorsque l'estimateur LS (B.2) est une valeur d'initialisation particulièrement mauvaise.

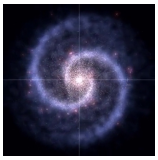
B.1.2 LTS adaptatif (ALTS)

Un choix classique pour le paramètre de rejet de la procédure LTS est $\alpha = 1/2$. Dans ce cas, l'estimateur LTS utilise la moitié des observations, celles ayant la variance estimée la plus faible. Cela donne un estimateur robuste avec un point de rupture de 50% similaire à la médiane par exemple. Toutefois l'efficacité asymptotique [1] de cet estimateur est de seulement 7% en l'absence de valeurs aberrantes (Rousseeuw et Leroy 1987, p.178-182). L'objectif d'une procédure LTS adaptative est alors d'estimer le nombre de valeurs non-aberrantes $\widehat{h}$ pour rester robuste tout en améliorant l'efficacité de l'estimateur. Notons respectivement h_0 et $\pi_0 = h_0/n$ le vrai nombre et la vraie proportion de valeurs non-aberrantes dans le vecteur d'observation $\boldsymbol{y}$. Considérons à présent la grandeur définie, pour $i = 1, \ldots, n$, par

$$s_i^2 = \frac{1}{i} \sum_{j=1}^{i} r_{(j)}^2, \tag{B.7}$$

où les résidus ordonnés sont obtenus par une première estimation par LTS avec $\alpha = 1/2$ (la dépendance des résidus en $\boldsymbol{\beta}$ est à présent omise par simplicité de notation). Il est aisé de remarquer que $(s_i^2)_{1 \leq i \leq n}$ est une séquence non-décroissante.

1. Pour un bruit symétrique, et en l'absence de valeurs aberrantes, l'efficacité asymptotique se ramène au ratio de la variance de l'estimateur LS divisé par la variance asymptotique de l'estimateur robuste.



B.1.2.1 Variance connue σ^2

Lorsque la variance du bruit σ^2 est connue, un estimateur $\widehat{h}$ du vrai nombre h_0 de valeurs non-aberrantes est donné par le critère d'arrêt suivant :

$$\widehat{h} = \max\left\{i \in \{1, \dots, n\} : \ s_i^2 \leq \sigma^2\right\}. \tag{B.8}$$

En effet, on fait l'hypothèse que les résidus que les résidus associés aux valeurs aberrantes sont plus grandes que les h_0 résidus associés aux valeurs non-aberrantes. Dans ce cas, la procédure LTS initiale avec $\alpha = 1/2$ donne un estimateur consistant de $\boldsymbol{\beta}$ sous réserve que $h_0 \geq \lceil n/2 \rceil$. De plus, lorsque $i = h_0$, s_i^2 est simplement la variance empirique des h_0 résidus non-aberrants. Cela donne un estimateur consistent de la variance σ^2. Inversement, lorsque $i > h_0$, s_i^2 est un estimateur biaisé par valeur supérieure de σ^2 avec un biais augmentant rapidement avec i puisque les résidus aberrants sont plus grands que les autres. Par conséquent, on s'attend à ce que $\widehat{h}$ soit un estimateur de h_0 plutôt conservatif.

Une nouvelle estimation par LTS peut alors être appliquée sur les $\widehat{h}$ plus petits résidus. Ces deux étapes :

- estimation $\widehat{h}$ de h_0 par (B.8),
- calcul de l'estimateur LTS résultant,

peuvent éventuellement être répétées jusqu'à convergence des estimateurs, comme décrit dans le cas général (variance inconnue) dans l'algorithme 8 décrit ci-après. En pratique, on peut toutefois arrêter l'algorithme au bout d'une ou deux itérations pour des gains en temps de calcul.

B.1.2.2 Variance inconnue

Dans le cas général, la variance σ^2 du bruit est inconnue. Une première estimation de cette variance peut être obtenue à l'aide de l'estimateur MAD (*median absolute deviation*). Rappelons que sous l'hypothèse de bruit blanc gaussien, l'estimateur de la variance à l'aide du MAD s'écrit :

$$\widehat{\sigma}_{\text{MAD}} = \frac{1}{\Phi^{-1}(3/4)} \left| r_{(\lceil n/2 \rceil)} \right|, \tag{B.9}$$

où Φ^{-1} est la fonction quantile de la loi normale. Cette valeur estimée peut alors être injecté dans l'équation (B.8) afin d'obtenir un premier estimateur $\widehat{h}$ de h_0. De plus, le paramètre de variance peut désormais être estimé par

$$\widehat{\sigma}^2 = \frac{1}{\widehat{h}} \sum_{i=1}^{\widehat{h}} r_{(i)}^2. \tag{B.10}$$

On obtient finalement l'algorithme itératif décrit dans Alg. 8 et appelé *adaptive least trimmed squares* (ALTS).

B.2 Application pour le filtrage de spectres

Une des difficultés majeures de l'étude des structures extragalactiques est que leurs raies d'émission sont souvent de faible intensité par rapport au continuum spectral (ou ligne de

Algorithme 8 Adaptive LTS (ALTS)

Entrées : $\boldsymbol{X}$ (matrice des régresseurs), $\boldsymbol{y}$ (données), $\boldsymbol{\beta}^0$ (valeur initiale)

Initialisation :
Calcul de l'estimateur $\boldsymbol{\beta}^1$ par LTS(X,$\boldsymbol{y}$,h^0,$\boldsymbol{\beta}^0$), avec $h^0 = \lceil n/2 \rceil$, où LTS fait référence à l'algorithme 7
Calcul de $\widehat{\sigma}_{MAD}$ par (B.9)
Calcul de h^1 en injectant $\widehat{\sigma}^2_{MAD}$ dans (B.8)

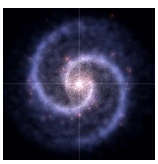
Boucle :
while $h^k \neq h^{k-1}$ **do**
    $\widehat{\boldsymbol{\beta}}^k = \text{LTS}(X, \boldsymbol{y}, \widehat{h}^{k-1}, \widehat{\boldsymbol{\beta}}^{k-1})$
    Calcul de $\widehat{\sigma}^{2,k}$ par (B.10)
    Calcul de h^k en injectant $\widehat{\sigma}^{2,k}$ dans (B.8)

Sortie : $\widehat{\boldsymbol{\beta}}^k$

base) des spectres des galaxies environnantes. Il est donc nécessaire de chercher à estimer puis supprimer ces lignes de base. Dans [Cleveland 1981], la méthode LOWESS est développée afin de traiter des données de plusieurs milliers de valeurs à l'aide de régressions locales pondérées. Dans le cadre du lissage de séries temporelles (ou de spectres), cela revient à utiliser une fenêtre glissante centrée en chaque échantillon et d'estimer la valeur lissée en ce point. Les échantillons au sein de la fenêtre glissante sont pondérés selon une fonction appelée *tricube* (voir [Cleveland 1981] pour une analyse du choix de cette fonction). Cette fonction est définie par $f(q) = (1-|q/l|^3)^3$ où $q \in \{-l, \ldots, l\}$ est la position relative au sein de la fenêtre de taille $2l+1$. Puisque cette régression paramétrique est effectuée localement sur chaque point, on obtient globalement un estimateur non paramétrique de la ligne de base. Cet estimateur n'est par défaut pas robuste, bien qu'il existe une version robuste, qui consiste à itérer le processus en appliquant une nouvelle pondération pénalisant les plus grands résidus [Cleveland 1981].

S'inspirant de la stratégie développée pour LOWESS, une régression locale à l'aide d'une fenêtre glissante pondérée est utilisée. Dans notre méthode, la régression locale est effectuée par l'algorithme ALTS décrit précédemment, avec une régression polynomiale d'ordre un. Chaque point du signal est ainsi estimé par une approche ALTS appliquée sur une fenêtre de voisinage pondérée. Un lissage est appliqué ensuite en post-traitement en utilisant à nouveau la fonction tricube. Il est à noter que cette procédure d'estimation ne dépend que d'un paramètre : la demi-largeur l de la fenêtre.

L'algorithme complet est décrit par l'algorithme 9, où $\boldsymbol{F}^{1/2} = \text{diag}\left(\{f(i)\}_{-l \leq i \leq l}\right)$ indique la matrice diagonale de la racine carrée des poids donnés par la fonction tricube, et $\boldsymbol{x}_k = (1, t_k)^T \in \mathbb{R}^2$ est la valeur du polynôme du premier ordre au temps t_k.



Algorithme 9 Estimation de ligne de base par ALTS

Entrées : $\boldsymbol{y}$ (données), l (demi-largeur de la fenêtre)

for i in (1, n) **do**
    Calcul de $X_i^l = \boldsymbol{F}^{1/2} \times [\boldsymbol{x}_{i-l}, ..., \boldsymbol{x}_{i+l}]^T$
    Calcul de $\boldsymbol{y}_i^l = \boldsymbol{F}^{1/2} \times (y_{i-l}, \ldots, y_{i+l})^T$
    Calcul de $\boldsymbol{\beta}_i = \text{ALTS}(X_i^l, \boldsymbol{y}_i^l, \boldsymbol{\beta}_{i-1})$ en utilisant Alg. 8
    Calcul de $\widehat{y}_i = (\boldsymbol{X}_i^l \boldsymbol{\beta}_i)_i$

Lissage de $\widehat{\boldsymbol{y}}$ à l'aide de régressions locales et d'une fonction tricube.

Sortie : $\widehat{\boldsymbol{y}}$

B.3 Simulation et résultats

B.3.1 Détection de valeurs aberrantes

On commence par tester l'algorithme 8 dans le cadre d'une régression polynomiale d'ordre trois. On construit pour cela une ligne de base polynomiale d'ordre trois, dégradée par un bruit gaussien et deux pics (de forme gaussienne) correspondants aux valeurs aberrantes. La procédure oracle consiste en une régression au sens des moindres carrés en utilisant uniquement les valeurs non-aberrantes.

La table B.1 affiche les résultats d'estimation de la ligne de base pour l'algorithme ALTS, avec connaissance ou non de la variance du bruit, pour l'algorithme α-LTS algorithm où $\alpha = 0.8$ et pour la procédure oracle, selon différentes valeurs de la proportion de valeurs non-aberrantes π_0. Tous les algorithmes de type LTS sont initialisés avec l'estimateur des moindres carrés. On peut voir que π_0 est correctement estimé par ALTS, avec et sans connaître la variance, dans le cas où $\pi_0 \geq 0.8$. Pour $\pi_0 = 0.7$ le nombre de valeurs aberrantes détectées est légèrement inférieur à la vraie valeur, même lorsque la variance est connue. Cela s'explique par la difficulté accrue de différencier la fin du pic et le bruit, comme illustré sur la figure B.1.

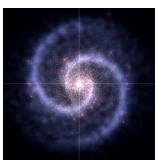
Lorsque le facteur de rejet α de LTS correspond à la vraie proportion π_0 de données non-aberrantes, la procédure LTS est légèrement plus efficace que l'approche adaptative. Dans tous les autres cas, elle est moins performante que la procédure ALTS en terme d'erreur quadratique moyenne intégrée le long du spectre (EQMI). On peut également remarquer que lorsque π_0 tend vers un, la procédure ALTS atteint l'efficacité de l'oracle. Ces résultats illustrent le fait que la procédure ALTS est à la fois robuste aux valeurs aberrantes et asymptotiquement efficace en l'absence de ces valeurs aberrantes.

B.3.2 Performance sur des spectres simulés

Un spectre de référence est construit à partir de données réelles MUSE. La ligne de base est construite à partir d'un moyennage spatial des spectres d'une galaxie dans les données MUSE, filtré ensuite par un filtre passe-bas récursif. Deux raies d'émission sont extraites de données réelles et ajoutées. Un bruit gaussien est ensuite ajouté, avec des paramètres estimés sur les données réelles. On peut voir une réalisation du signal résultant sur la figure B.2 (a).

	ALTS	ALTS (σ^2 connu)	LTS ($\alpha = 0.8$)	Oracle
π_0 = *0.7*				
$\hat{\pi}_0$	0.76	0.73	(0.8)	(0.7)
EQMI	0.43	0.36	0.68	0.13
Variance	0.20	0.20	0.17	0.13
Bias^2	0.23	0.16	0.51	1×10^{-5}
π_0 = *0.8*				
$\hat{\pi}_0$	0.81	0.80	(0.8)	(0.8)
EQMI	0.21	0.21	0.18	0.15
Variance	0.18	0.18	0.17	0.15
Bias^2	0.025	0.032	0.015	1×10^{-4}
π_0 = *0.9*				
$\hat{\pi}_0$	0.89	0.89	(0.8)	(0.9)
EQMI	0.11	0.12	0.18	0.10
Variance	0.11	0.11	0.18	0.10
Bias^2	1.4×10^{-4}	1×10^{-3}	5×10^{-4}	6×10^{-5}
π_0 = *1*				
$\hat{\pi}_0$	0.98	0.98	(0.8)	(1)
EQMI	0.09	0.09	0.20	0.087
Variance	0.09	0.09	0.20	0.087
Bias^2	7.4×10^{-5}	9.6×10^{-5}	9.8×10^{-5}	9.6×10^{-5}

TABLE B.1 – Performances des algorithmes ALTS, ALTS avec σ^2 connu, α-LTS avec $\alpha = 0.8$ et oracle pour π_0 de 0.7 à 1. Les valeurs sont intégrées sur le long du signal et calculées sur 1000 simulations Monte Carlo



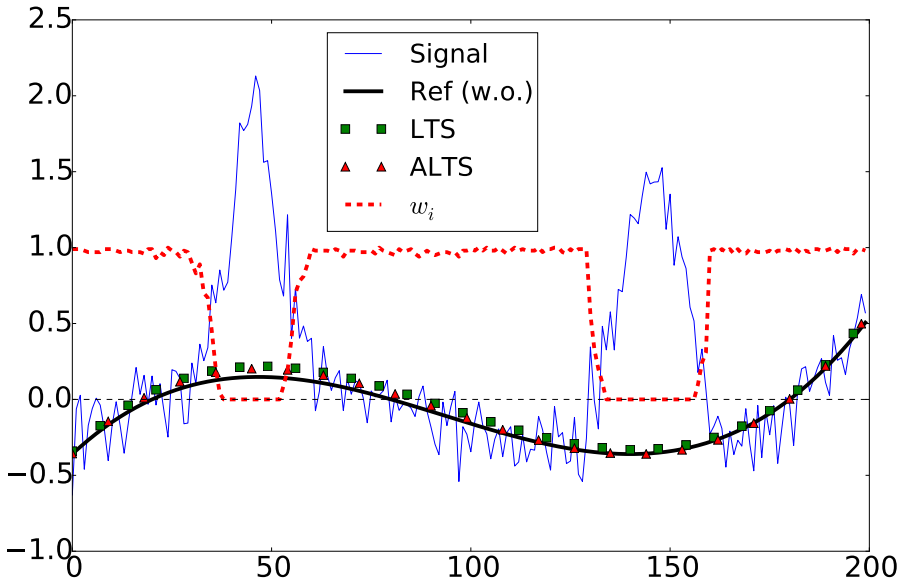


FIGURE B.1 – Estimation moyenne de la ligne de base sur 1000 simulations Monte-Carlo avec $\pi_0 = 0.7$. L'estimation par ALTS est indiquée par des marqueurs ▲, l'estimation par α-LTS avec $\alpha = 0.8$ est indiquée par des marqueurs ■, une réalisation du signal bruité est affichée par une courbe bleue, la vraie ligne de base est indiqué en trait noir épais, et les poids moyens w_i (détection de valeurs non-aberrantes) en pointillés rouges.

La méthode proposée est comparée avec l'algorithme BEADS et la version robuste de LOWESS. Les paramètres pour ces deux méthodes ont été choisis les plus performants possibles. La fréquence de coupure de BEADS a été dans un premier temps fixée à la même valeur que le paramètre du filtre passe-bas utilisé pour construire la ligne de base de référence. Après essais, les meilleurs résultats étaient toutefois optimaux pour une fréquence de coupure légèrement plus faible. Les autres paramètres ont également été optimisés par essai-erreur. La taille de la fenêtre est identique dans les procédures LOWESS et ALTS.

	ALTS	LOWESS	BEADS
Spectre entier			
EQMI	3.76	4.20	5.84
Variance	2.98	3.26	2.59
Biais^2	0.78	0.94	3.25
Zone de la raie			
EQMI	0.26	0.45	0.91

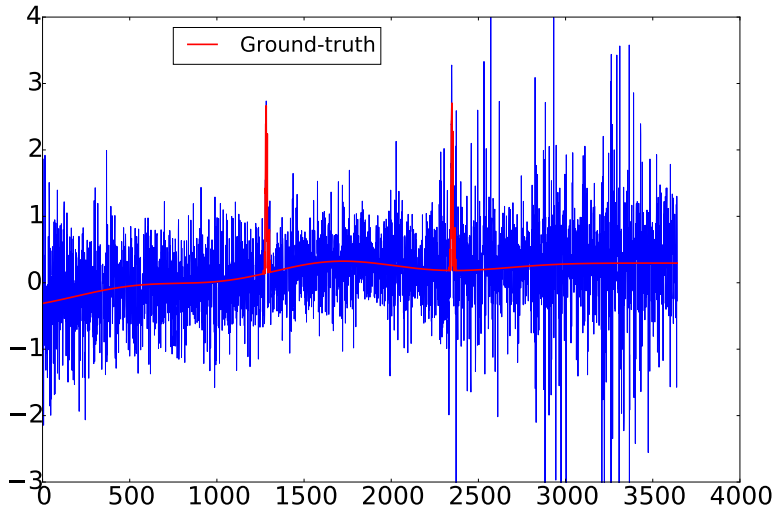
TABLE B.2 – Comparaison des procédures ALTS, LOWESS et BEADS. Les valeurs sont intégrées soit sur tout le signal, soit sur une fenêtre de taille 200 centrée sur le pic d'une raie.

Les résultats présentés dans la table B.2 sont obtenus à partir de 200 simulations Monte-Carlo. Les résultats de la procédure ALTS semblent légèrement meilleurs que les approches de

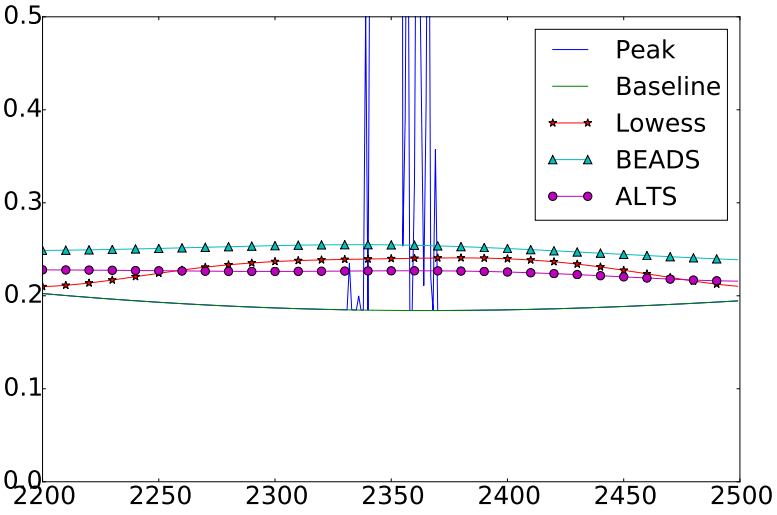
l'état de l'art, BEADS et LOWESS, en particulier dans la zone de la raie d'émission, comme illustré par la figure B.2(b).

B.3.3 Paramètres et coût calculatoire

L'algorithme final ne dépend que d'un seul paramètre : la taille de la fenêtre. Cette dernière impacte principalement la variance de l'estimation : plus elle est grande, plus l'estimation est lissée. Sur données expérimentales, il semble que son réglage soit relativement peu sensible. Le coût calculatoire est néanmoins relativement élevé : quelques secondes pour un signal de 3600 valeurs dans son implémentation Python, contre 200ms pour l'implémentation MATLAB de BEADS et une seconde pour la version Python de LOWESS. Il est à noter toutefois que BEADS dépend de quatre paramètres : la fréquence de coupure et trois poids de pénalisations qui doivent être réglé avec précision (notamment la fréquence de coupure).



(a) Réalisation d'un signal à partir de données MUSE (courbe rouge), qui est bruité ensuite (courbe bleue).



(b) Estimation de la ligne de fond autour du second pic d'émission. Le signal complet est en bleu, la ligne de fond est en vert, l'estimation par ALTS est représenté par des ●, l'estimation par LOWESS est représenté par des ∗, l'estimation par BEADS est représenté par des ▲.

FIGURE B.2 – Estimation de la ligne de fond sur données simulées pour les procédures ALTS, LOWESS, et BEADS, moyennée sur 200 réalisations.

Annexe C

Approche par classification pour la détection de halos

C.1 Méthode

L'approche choisie ici afin de répondre au problème de détection des halos galactiques se fonde sur des approches de *machine learning* classiques. Dans une bande spectrale étroite autour de l'émission Lyman, la galaxie étudiée et le potentiel halo voisin sont supposés partager une signature spectrale assez similaire (raie d'émission), alors que les autres pixels contiennent soit uniquement du bruit soit le spectre de galaxies voisines au contenu a priori très différent dans cette bande spectrale. En effectuant une classification à deux classes dans une bande spectrale étroite autour de la raie, l'objectif est d'obtenir une classe contenant tous les pixels liés à une émission Lyman-α et une classe regroupant les autres pixels.

C.1.1 Pré-traitements

Réduction des données

Avant tout autre traitement du cube de données, une réduction des données par le cube de variance est effectué, afin notamment de diminuer l'influence des pixels anormaux.

Pour chaque valeur du spectre d'un pixel (voxel) $x_{i,\lambda}$, où i est la position spatiale et λ la coordonnée spectrale, on obtient le voxel réduit $\tilde{x}_{i,\lambda}$ en normalisant par la variance du voxel $v_{i,\lambda}$:

$$\tilde{x}_{i,\lambda} = \frac{x_{i,\lambda}}{\sqrt{v_{i,\lambda}}}$$

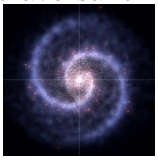
Chaque voxel est donc désormais la réalisation d'une variable aléatoire de variance unitaire.

Soustraction du continuum spectral

Afin de réduire les sources de nuisances dues notamment aux galaxies voisines, on effectue une soustraction des lignes de base des spectres. Cette soustraction peut se faire soit par un filtre médian soit à l'aide de la méthode proposée dans [Bacher et al. 2016a].

Filtrage adapté

Afin d'améliorer le SNR, un filtrage adapté à la PSF de l'ensemble {instrument+atmosphère} peut être appliqué. La PSF est supposée séparable en sa composante spatiale, la FSF et sa composante spectrale LSF. Par ailleurs, la composante de l'instrument seul est négligeable devant la composante due aux turbulences de l'atmosphère. L'hypothèse clé ici est que la majeure partie du bruit est due à l'instrument alors que la PSF s'applique principalement au signal traversant



l'atmosphère avant d'atteindre l'instrument. Appliquer cette PSF permet donc d'améliorer significativement le RSB des structures galactiques. L'inconvénient majeure de l'utilisation de la composante spatiale de la PSF dans le filtrage adaptée est l'étalement spatial engendré, qui peut perturber l'interprétation finale.

C.1.2 Classification non supervisée

Une zone spatiale et spectrale est définie autour de chaque galaxie (centrée spatialement sur le centre de la galaxie et spectralement sur le pic d'émission), le spectre de chaque pixel sur cette bande spectrale étant choisi comme vecteur de caractéristiques.

Choix d'une métrique

Les différentes métriques explorées sont la distance euclidienne, la distance spectrale angulaire (SAD) et la *spectral information divergence* (SID, Chang 1999), définies comme suit : Pour $\mathbf{x} = x_1...x_n$, et $\mathbf{y} = y_1, ...y_n$

- $d_{EUC} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
- $d_{SAD} = cos^{-1}\left(\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}\right)$
- $d_{SID} = \sum_{i=1}^{n} x_i \times log(\frac{x_i}{y_i}) + y_i \times log(\frac{y_i}{x_i})$ avec $\forall i, x_i > 0$ et $y_i > 0$

Les histogrammes des différentes distances entre les spectres des pixels de la zone étudiée et un spectre de référence sont montrés sur la figure C.1. Le spectre de référence est estimé au centre de la galaxie. Après étude sur les données réelles, la métrique SAD, associée à une approche de *spectral clustering*, a été retenue. Le SAD est moins sensible à l'intensité totale du spectre qu'à la distribution d'intensité sur la bande spectrale considéré. Il permet ainsi d'être robuste face aux fortes dynamiques présentes dans les données, et de favoriser l'association de signatures spectrales proches même lorsqu'elles sont de faible intensité. Notons que le SID se révèle un bon candidat sur données simulées mais se limite à des signaux à valeurs strictement positives. Les spectres issus des données réelles considérées ici sont trop bruités pour vérifier cette contrainte, notamment une fois la réduction des données et la soustraction des lignes de base appliquées.

Spectral clustering

Le spectral clustering (Ng et al. 2002) s'appuie sur une matrice d'affinité associée à un graphe pondéré. Ici, un noeud $\boldsymbol{x}_i$ est un spectre de λ bandes spectrales, associé à un pixel d'une image centrée sur la galaxie étudiée.

Notons $\boldsymbol{M}$ une matrice de distance associée au graphe, définie par : pour chaque noeud $\boldsymbol{x}_i$ et $\boldsymbol{x}_j$ du graphe, $M_{i,j} = d(x_i, x_j)$. Une matrice de similarité $\boldsymbol{S}$ peut alors être dérivée de la matrice de distance par $S_{ij} = \exp^{-\frac{M_{ij}^2}{\sigma^2}}$, avec σ un paramètre clé correspondant à la taille caractéristique d'un cluster. La matrice du Laplacien normalisé $\boldsymbol{L}$ du graphe est alors définie par

$$\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{S}\boldsymbol{D}^{-\frac{1}{2}}$$

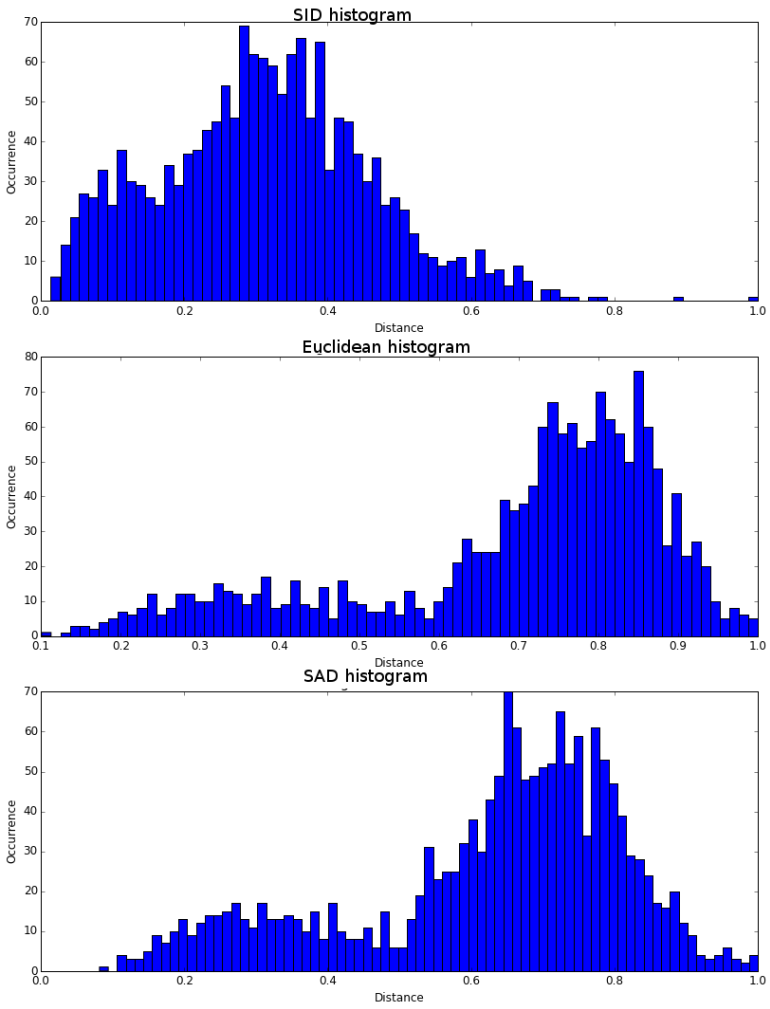
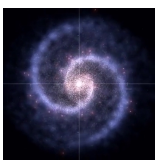


FIGURE C.1 – Comparaisons des histogrammes pour les différentes distances ((normalisées entre 0 et 1) entre les spectres du voisinage d'une galaxie et le spectre de référence du centre de la galaxie.



avec $\boldsymbol{D}$ la matrice de degré : $D_{ii} = \sum_{j=1}^{n} S_{ij}$.

Le spectral clustering consiste alors à trouver les k premiers vecteurs propres du Laplacien, k étant le nombre de clusters attendus, et à projeter les données sur l'espace engendré par ces vecteurs propres. N'importe quel algorithme de classification non supervisé peut ensuite être utilisé sur les données réduites. Ici, on utilisera un algorithme K-means pour identifier les k clusters.

Estimation de σ

Comme mentionné précédemment, le choix du paramètre σ utilisé pour construire la matrice de similarité est crucial. De nombreuses stratégies peuvent être essayées pour estimer automatiquement σ. Minimiser le ratio entre la variance intra-classe et la variance extra-classe est une approche classique mais elle conduit ici à donner une importance trop grande aux pixels du coeur de la galaxie, au RSB bien plus significatif que les pixels du halo qui sont alors rejetés.

En revenant sur l'interprétation physique de σ comme étant le paramètre réglant taille caractéristique d'une classe, il a été choisi de le régler à partir de la distance moyenne intra-classe. Comme on peut voir sur l'histogramme des distances SAD montré sur la figure C.1, l'histogramme peut être approximativement décomposé en un mélange de deux distributions. La première est une distribution gaussienne des distances entre le spectre de référence et un spectre de fond "aléatoire" et la seconde est une distribution de moyenne bien plus faible, qui correspond à un ensemble de données très proches les unes des autres (faible valeurs de SAD), et donc *a priori* à l'ensemble des pixels recherchés.

L'estimation des paramètres de la première composante gaussienne est faite par un algorithme d'espérance-maximisation (EM), dont la première étape consiste à estimer la probabilité pour un pixel d'appartenir au processus gaussien, et la seconde à estimer les paramètres gaussiens. Bien que cet algorithme réalise déjà une opération de classification, il est à noter que ses performances sont plutôt médiocres. Ce résultat est toutefois conservé afin de régler les paramètres de l'approche de spectral clustering, bien plus efficace. Bien que l'histogramme représente la distance à un spectre de référence et non les distances intra-groupes, il semble raisonnable de considérer que la distance interne de la classe "bruit" est de l'ordre de grandeur de l'écart-type estimé sur le mode gaussien détecté. Le paramètre σ est donc fixé à cette dernière valeur.

Il est à noter qu'une approche semi-supervisée basée sur les travaux de [Wang et Davidson 2010] a également été exploré, afin de tenter de contourner le problème d'estimation de σ. L'idée est d'ajouter un certain nombre de contraintes sur des paires de pixels connues pour appartenir ou non au même groupe. Bien que cette approche semble intéressante puisqu'elle permet l'injection d'informations connues, en pratique les résultats obtenus n'étaient pas plus satisfaisants que l'approche non-supervisée, car une proportion extrêmement importante de pixels nécessitait d'être labellisée afin d'obtenir une classification satisfaisante.

C.1.3 Post-traitement

Afin d'améliorer la robustesse de la classification, une régularisation spatiale à l'aide d'un modèle de Potts est proposée, implémentée à l'aide d'une approche par recuit simulé. Le modèle de Potts (Pony et al. 2000) implique deux contributions d'énergies ; l'une (U_{ext}) est liée aux interactions entre pixels et est contrôlée par un paramètre de réglage β qui sert de "température",

et la seconde (U_{int}) est l'énergie propre associée à chaque pixel. On cherche alors à trouver la configuration de labels c des pixels étudiés qui minimise l'énergie totale U ou de façon équivalente qui maximise la probabilité $P(C = c) = \frac{e^{-U(c)}}{Z}$ avec Z le terme de normalisation.

Dans notre contexte, pour chaque pixel, l'énergie interne est donnée par la distance euclidienne entre le pixel et le centre de sa classe, et le terme de régularisation (ou d'interaction) est le nombre de voisins (par exemple dans un 8-voisinage) qui ne possèdent pas la même classe.

Notons N le nombre de pixels, $\{c_i\}_{i=1..N}$, la configuration des labels $\{l_i\}_{i=1..N}$ des pixels $\{i\}_{i=1..N}$ résultant de la classification et k le nombre de clusters (2 ici). Alors pour $j \in [\![0, k]\!]$,

$$P(l_i = j|c_i) = \frac{e^{-(U_{int}^{\setminus i} + \beta U_{ext}^{\setminus i} + U_{l_i=j}^{i})}}{Z}$$

avec $U_{...}^{\setminus i}$ le terme d'énergie sommé sur tous les pixels sauf i et $U_{l_i=j}^{i}$ l'énergie du pixel i labellisé dans la classe j.

Puisque le choix du paramètre de régularisation β est un problème délicat, une marginalisation de β est proposée, en supposant une loi *a priori* Gamma non informative pour β désormais vu comme une variable aléatoire. Afin de réaliser cette marginalisation, Z, désormais fonction de β, est approchée par $Z(\beta) \simeq \beta^{-1}$. Il vient alors que la probabilité locale, conditionnellement à β peut être approchée par

$$P(c = l_i|\beta) = \begin{cases} \beta e^{-\beta U_{ext}(c=l_i)} \text{ si } U_{ext}(c = l_i) \neq 0 \\ \beta e^{-\frac{\beta}{2}} \text{ sinon} \end{cases}$$

En utilisant un *a priori* Gamma de paramètres α_0,α_1 (fixés à 1 ici), pour β, et la formule de Bayes, β peut être marginalisé. On obtient :

$$P(c = l_i|C) \propto e^{-\frac{U_{int}(C,c=l_i)}{T}} \left(\sum_{c,U_{ext}(c=l_i)\neq 0} U_{ext}(c = l_i) + \frac{N_0}{2} + \frac{1}{\alpha_1} \right)^{-\frac{N+\alpha_0}{T}}$$

avec T la température du recuit, N le nombre de pixels et N_0 le nombre de cliques homogènes, i.e. où $U_{ext}(c = l_i) = 0$.

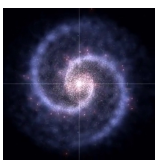
La minimisation de l'énergie totale sur tous les pixels est alors obtenue par un algorithme de recuit simulé utilisant cette formulation de la probabilité *a posteriori* de configuration.

Le résultat de la régularisation utilisant un β marginalisé est montré sur la figure C.2.

C.2 Résultats

C.2.1 Données MUSE

Le champ *Hubble Deep Field South* (HDFS), montré sur la figure C.3 est un champ les plus profonds jamais observé de l'Univers par le télescope spatial Hubble et a ainsi été une des premières région du ciel visées par MUSE. Par la suite on appellera également par abus de langage HDFS l'observation MUSE de ce champ. Les détails de la réduction des données de l'observation MUSE du HDFS sont données dans [Bacon et al. 2015], ainsi que les premières analyses spectrales des galaxies détectées. Afin de tester cette approche de détection de halo par classification, plusieurs objets d'intérêt ont été sélectionnés dans les données HDFS, certains avec un halo attendu, d'autres sans. Des zones 3D de 40 par 40 pixels par 30 bandes spectrales sont sélectionnées, centrées spatialement sur l'objet d'intérêt et spectralement à la position du pic de la raie Lyman-α.



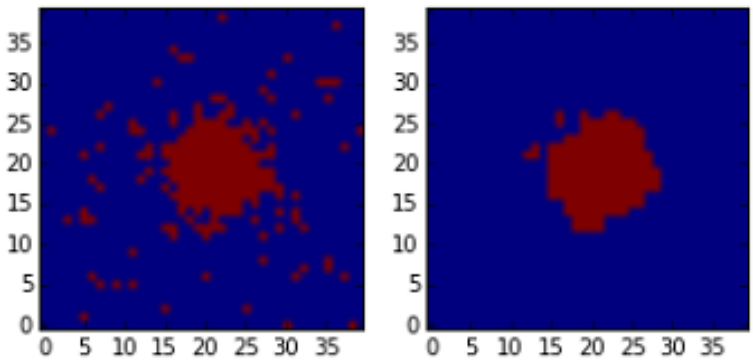


FIGURE C.2 – Classification d'un ensemble galaxie+halo sans et avec une régularisation (où β est marginalisé)

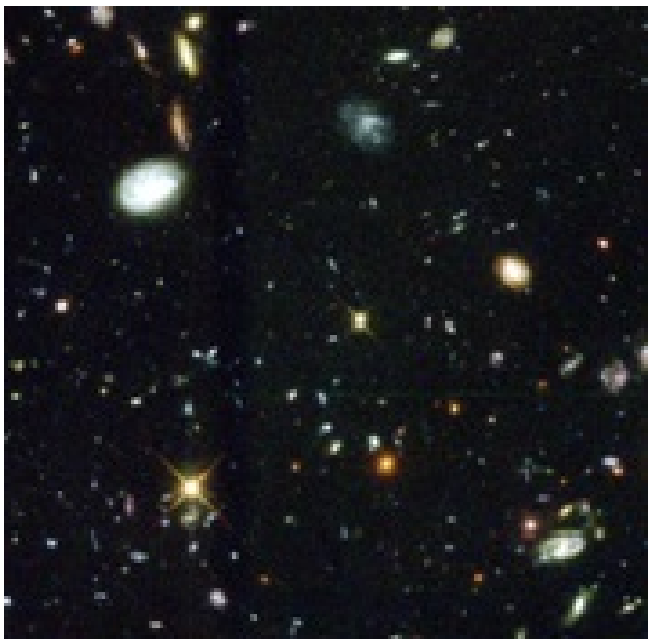


FIGURE C.3 – Image HDFS

C.2.2 Chaîne de traitement

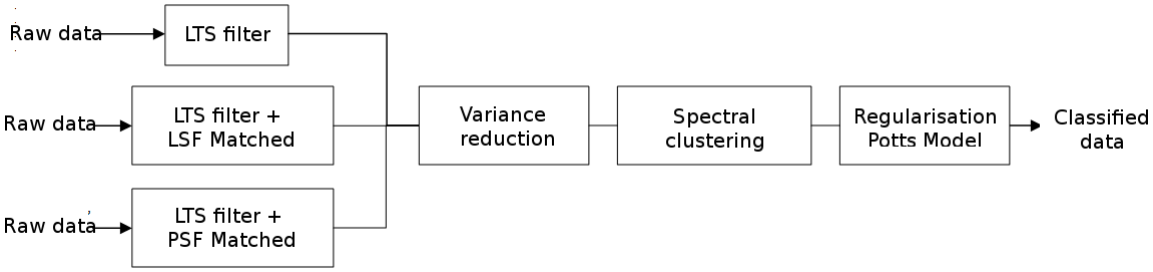


FIGURE C.4 – Résumé de la chaîne de traitement, à partir du cube de données produit par la chaîne de réduction de données du consortium.

L'estimation et la soustraction du continuum sont effectués par la méthode ALTS développée dans [BACHER et al. 2016a] à l'aide d'une fenêtre glissante de taille 100. Trois cas sont alors considérés : sans filtre adapté, avec un filtrage adapté à la LSF, ou avec un filtrage adapté à la PSF complète (LSF et FSF). Les étapes de classification sont ensuite identiques pour les trois options, basées sur le spectral clustering décrit précédemment. Enfin une régularisation spatiale par modèle de Potts est appliquée, implémentée en marginalisant le paramètre β.

C.2.3 Analyse

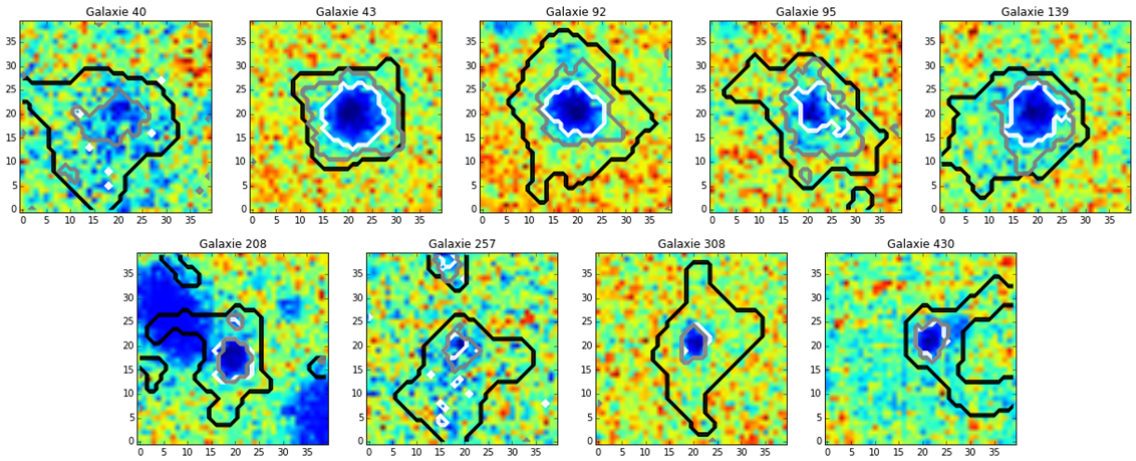


FIGURE C.5 – Contour de classification pour chaque chaîne de traitement. Les résultats sans filtre adapté sont en blanc, ceux obtenus en utilisant le filtrage spectral par la LSF sont en gris et ceux obtenus avec le filtrage par la PSF complète sont en noir. La distance SAD au centre de la galaxie (avant tout pré-traitement) est indiquée par dessous (bleu signifie proche de 0 et indiquant ainsi un spectre proche (au sens du SAD) du spectre central. Les ID des galaxies font référence au catalogue de [BACON et al. 2015]

Les résultats des trois différentes chaînes de traitement sur les 9 galaxies choisies dans le HDFS sont montrés sur la figure C.5. Le résultat le plus notable est que l'utilisation du filtrage adapté à la PSF complète (spatiale et spectral) donne clairement les détections les plus étendues. Comme souligné précédemment, ce résultat est bien sûr biaisé par l'effet d'étalement spatial

du filtrage par la FSF. Néanmoins, il permet de révéler des structures asymétriques atypiques dans le voisinage de plusieurs galaxies.

L'importance de l'étape de soustraction du continu est souligné par exemple sur le 6ème objet : on peut voir que la galaxie d'intérêt est très proche spatialement d'autres galaxies qui apparaissent similaires sous la métrique SAD. La mesure de similarité SAD n'est en effet pas suffisamment discriminante pour différencier la raie Lyman de la galaxie d'intérêt de la très forte émission continue de ces galaxies. Les pré-traitements permettent de soustraire ces galaxies du cluster recherché. Par conséquence, seule l'extension liée à la galaxie d'intérêt semble être détectée.

C.3 Conclusion

Ces approches préliminaires montrent bien que certaines galaxies sont très certainement entourées par un halo ; des structures très asymétriques sont ainsi révélées autour de ces galaxies.

Afin d'améliorer l'estimation du support spatial du halo, il semble nécessaire de prendre en compte la variabilité spectrale du halo ciblé ; en effet la signature spectrale de la source étendue peut varier lorsqu'on s'éloigne du centre de la galaxie. En l'état, des approches de classification directes comme celle étudié ici ne sont peut-être pas les plus pertinentes pour aborder ce problème. En particulier, il semble assez difficile d'obtenir une méthode de classification complètement automatique pour ce problème : dans notre approche, le paramètre σ de la matrice de similarité devrait par exemple être réglé de façon adaptative pour prendre en compte la variabilité spectrale. De plus, les contraintes classiques ajoutées dans le cadre d'une approche semi-supervisée n'ont étonnamment pas amélioré les résultats, ce qui n'a pas pu être expliqué de façon satisfaisante. D'autres approches seraient donc à explorer pour injecter de façon plus efficace des connaissances *a priori*. Enfin, la limite majeure de ce type d'approche est la difficulté d'assurer un contrôle fiable des erreurs de détection. On voit en effet que l'utilisation de la FSF comme filtre adapté semble nécessaire pour observer des extensions significatives, mais ce filtrage adapté introduit un biais (en étalant artificiellement les extensions spatiales) qu'il est difficile d'estimer et de contrôler. C'est la raison pour laquelle ces travaux sont restés pour l'instant exploratoire et que la problématique de détection s'est recentré sur des approches par test d'hypothèses comme présenté dans la partie II de ce manuscrit.

Bibliographie

Aharon, Michal, Michael Elad et Alfred Bruckstein (2006). « K-SVD : An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation ». In : *IEEE Transactions on signal processing* 54.11, p. 4311–4322 (cf. p. 69).

Aho, Ken, DeWayne Derryberry et Teri Peterson (2014). « Model selection for ecologists : the worldviews of AIC and BIC ». In : *Ecology* 95.3, p. 631–636 (cf. p. 45).

Akaike, Hirotugu (1974). « A new look at the statistical model identification ». In : *IEEE transactions on automatic control* 19.6, p. 716–723 (cf. p. 38).

Akaike, Htrotugu (1973). « Maximum likelihood identification of Gaussian autoregressive moving average models ». In : *Biometrika* 60.2, p. 255–265 (cf. p. 45).

Akhlaghi, Mohammad et Takashi Ichikawa (2015). « Noise-based Detection and Segmentation of Nebulous Objects ». In : *The Astrophysical Journal Supplement Series* 220.1, p. 1 (cf. p. 62).

Arias-Castro, Ery, Emmanuel J Candès et Yaniv Plan (2011). « Global testing under sparse alternatives : ANOVA, multiple comparisons and the higher criticism ». In : *The Annals of Statistics*, p. 2533–2556 (cf. p. 83, 85, 87, 125).

Arlot, Sylvain, Alain Celisse et al. (2010). « A survey of cross-validation procedures for model selection ». In : *Statistics surveys* 4, p. 40–79 (cf. p. 50).

Bacher, Raphael, Florent Chatelain et Olivier Michel (2016a). « An adaptive robust regression method : application to galaxy spectrum baseline estimation ». In : *IEEE International Conference on Accoustic, Speech and Signal Processing 2016* (cf. p. xviii, xix, 119, 135, 145, 151).

— (2016b). « Source Halo Advanced Detection and Estimation : une méthode de détection du Circum-Galactic Medium ». In : *Colloque du Groupe Hyperspectral de la Société Française de Photogrammétrie et de Télédétection (SFPT) 2016* (cf. p. xix).

Bacher, Raphael, Pierre Maho, Florent Chatelain et Olivier Michel (2016c). « Tracking the Lyman alpha emission line in the CircumGalactic Medium in MUSE data ». In : *EAS Publications Series* 78, p. 233–245 (cf. p. xx).

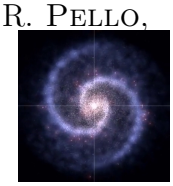
Bacher, Raphael, Florent Chatelain et Olivier Michel (2017a). « Détection de cibles spatialement structurées sous contrôle global d'erreur ». In : *Colloque du Groupe de recherche et d'étude de traitement du signal et des images (GRETSI) 2017* (cf. p. xix).

— (2017b). « Global error control procedure for spatially structured targets ». In : *IEEE European Signal Processing Conference 2017* (cf. p. xviii, xix).

Bacher, Raphael, Céline Meillier, Florent Chatelain et Olivier Michel (2017c). « Robust Control of Varying Weak Hyperspectral Target Detection With Sparse Nonnegative Representation ». In : *IEEE Transactions on Signal Processing* 65.13, p. 3538–3550 (cf. p. xviii, xix).

Bacon, R, J Brinchmann, J Richard, T Contini, A Drake, M Franx, S Tacchella, J Vernet, L Wisotzki, J Blaizot et al. (2015). « The MUSE 3D view of the Hubble Deep Field South ». In : *A&A* 575, A75 (cf. p. 3, 149, 151).

Bacon, R., S. Conseil, D. Mary, J. Brinchmann, M. Shepherd, M. Akhlaghi, P. Weilbacher, L. Piqueras, L. Wisotzki, D. Lagattuta, B. Epinat, A. Guerou, H. Inami, S. Cantalupo, C. Clastres, J.-B. Courbot, T. Contini, J. Richard, M. Maseda, R. Bouwens, N. Bouch'e, W. Kollatschny, J. Schaye, R. Anna Marino, R. Pello,



C. Herenz, B. Guiderdoni et M. Carollo (2017). « The MUSE Hubble Ultra Deep Field Survey : I. Survey description, data reduction and source detection ». In : *Astronomy and Astrophysics* (cf. p. 3, 4, 7, 59, 117, 121).

Bacon, Roland, Laure Piqueras, Simon Conseil, Johan Richard et Martin Shepherd (2016). « MPDAF : MUSE Python Data Analysis Framework ». In : *Astrophysics Source Code Library* (cf. p. 128).

Barber, Rina Foygel, Emmanuel J Candès et al. (2015). « Controlling the false discovery rate via knockoffs ». In : *The Annals of Statistics* 43.5, p. 2055–2085 (cf. p. 83, 101–103).

Barber, Rina Foygel et Aaditya Ramdas (2015). « The p-filter : multi-layer FDR control for grouped hypotheses ». In : *arXiv preprint arXiv :1512.03397* (cf. p. 83, 98).

Bayliss, Jessica D, J Anthony Gualtieri et Robert F Cromp (1998). « Analyzing hyperspectral data with independent component analysis ». In : *26th AIPR Workshop : Exploiting New Image Sources and Sensors*. International Society for Optics et Photonics, p. 133–143 (cf. p. 24).

Belloni, Alexandre et Victor Chernozhukov (2009). « Least squares after model selection in high-dimensional sparse models ». In : (cf. p. 49).

Benjamini, Yoav et Yosef Hochberg (1995). « Controlling the false discovery rate : a practical and powerful approach to multiple testing ». In : *Journal of the Royal Statistical Society. Series B (Methodological)*, p. 289–300 (cf. p. 82, 85, 94).

Benjamini, Yoav et Daniel Yekutieli (2001). « The control of the false discovery rate in multiple testing under dependency ». In : *Annals of statistics*, p. 1165–1188 (cf. p. 94).

Bertin, Emmanuel et Stephane Arnouts (1996). « SExtractor : Software for source extraction ». In : *Astronomy and Astrophysics Supplement Series* 117.2, p. 393–404 (cf. p. 9).

Bourguignon, Sébastien, David Mary et Éric Slezak (2011). « Restoration of astrophysical spectra with sparsity constraints : Models and algorithms ». In : *IEEE Journal of Selected Topics in Signal Processing* 5.5, p. 1002–1013 (cf. p. 79).

— (2012). « Processing MUSE hyperspectral data : Denoising, deconvolution and detection of astrophysical sources ». In : *Statistical Methodology* 9.1, p. 32–43 (cf. p. 9).

Breiman, Leo, Jerome Friedman, Charles J Stone et Richard A Olshen (1984). *Classification and regression trees*. CRC press (cf. p. 49, 51).

Candès, Emmanuel, Yingying Fan, Lucas Janson et Jinchi Lv (2016). « Panning for Gold : Model-free Knockoffs for High-dimensional Controlled Variable Selection ». In : *arXiv preprint arXiv :1610.02351* (cf. p. 101).

Carfantan, Hervé (2014). « Modèles, estimateurs et algorithmes pour quelques problèmes inverses de traitement du signal et d'images en sciences de l'univers ». Habilitation à Diriger des Recherches (HDR) (cf. p. 6, 29).

Chang, Chein-I (1999). « Spectral information divergence for hyperspectral image analysis ». In : *Geoscience and Remote Sensing Symposium, 1999. IGARSS'99 Proceedings. IEEE 1999 International*. T. 1. IEEE, p. 509–511 (cf. p. 118, 146).

Chen, Yi, Nasser M Nasrabadi et Trac D Tran (2011). « Sparse representation for target detection in hyperspectral imagery ». In : *IEEE Journal of Selected Topics in Signal Processing* 5.3, p. 629–640 (cf. p. 79).

Chiu, I, Shantanu Desai et Jiayi Liu (2016). « ComEst : A completeness estimator of source extraction on astronomical imaging ». In : *Astronomy and Computing* 16, p. 79–87 (cf. p. 106).

Cleveland, William S (1981). « LOWESS : A program for smoothing scatterplots by robust locally weighted regression ». In : *American Statistician*, p. 54–54 (cf. p. 135, 139).

Comon, Pierre (1994). « Independent component analysis, a new concept ? » In : *Signal processing* 36.3, p. 287–314 (cf. p. 24).

Courbot, Jean-Baptiste (2017). « Détection de sources quasi-ponctuelles dans des champs de données massifs ». Thèse de doct. Université de Strasbourg (cf. p. 8, 79).

Courbot, Jean-Baptiste, Emmanuel Monfrini, Vincent Mazet et Christophe Collet (2016). « Oriented Triplet Markov Field for Hyperspectral Image Segmentation ». In : *IEEE Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing* (cf. p. 79, 119, 120, 125).

— (2017a). « Arbres de markov triplets pour la segmentation d'images ». In : *Colloque du Groupe de recherche et d'étude de traitement du signal et des images (GRETSI) 2017* (cf. p. 79).

Courbot, Jean-Baptiste, Vincent Mazet, Emmanuel Monfrini et Christophe Collet (2017b). « Extended faint source detection in astronomical hyperspectral images ». In : *Signal Processing* 135, p. 274–283 (cf. p. 79, 97).

Donoho, David et Jiashun Jin (2004). « Higher criticism for detecting sparse heterogeneous mixtures ». In : *The Annals of Statistics* 32.3, p. 962–994 (cf. p. 83, 87).

— (2015). « Higher criticism for large-scale inference, especially for rare and weak effects ». In : *Statistical Science* 30.1, p. 1–25 (cf. p. 83).

Dunn, Olive Jean (1961). « Multiple comparisons among means ». In : *Journal of the American Statistical Association* 56.293, p. 52–64 (cf. p. 82).

Efron, Bradley (2012). *Large-scale inference : empirical Bayes methods for estimation, testing, and prediction*. T. 1. Cambridge University Press (cf. p. 89).

Efron, Bradley, Trevor Hastie, Iain Johnstone, Robert Tibshirani et al. (2004). « Least angle regression ». In : *The Annals of statistics* 32.2, p. 407–499 (cf. p. 44).

Esary, James D, Frank Proschan, David W Walkup et al. (1967). « Association of random variables, with applications ». In : *The Annals of Mathematical Statistics* 38.5, p. 1466–1474 (cf. p. 134).

Friedman, Jerome, Trevor Hastie et Rob Tibshirani (2010). « Regularization paths for generalized linear models via coordinate descent ». In : *Journal of statistical software* 33.1, p. 1 (cf. p. 44).

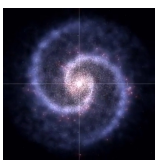
Fruchter, A et al. (2009). « HST multidrizzle handbook ». In : *HST MultiDrizzle, HST Data Handbooks* 1 (cf. p. 4).

Galatsanos, Nikolas P et Aggelos K Katsaggelos (1992). « Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation ». In : *IEEE Transactions on image processing* 1.3, p. 322–336 (cf. p. 50).

Geisser, Seymour (1975). « The predictive sample reuse method with applications ». In : *Journal of the American statistical Association* 70.350, p. 320–328 (cf. p. 50).

Genovese, Christopher R, Nicole A Lazar et Thomas Nichols (2002). « Thresholding of statistical maps in functional neuroimaging using the false discovery rate ». In : *Neuroimage* 15.4, p. 870–878 (cf. p. 81).

Genz, Alan et Frank Bretz (2009). *Computation of multivariate normal and t probabilities*. T. 195. Springer Science & Business Media (cf. p. 114, 134).



Golub, Gene H, Michael Heath et Grace Wahba (1979). « Generalized cross-validation as a method for choosing a good ridge parameter ». In : *Technometrics* 21.2, p. 215–223 (cf. p. 51).

Hall, Peter, Jiashun Jin et al. (2010). « Innovated higher criticism for detecting sparse signals in correlated noise ». In : *The Annals of Statistics* 38.3, p. 1686–1732 (cf. p. 83).

Haughton, Dominique MA et al. (1988). « On the choice of a model to fit data from an exponential family ». In : *The Annals of Statistics* 16.1, p. 342–355 (cf. p. 45).

Hoerl, Arthur E et Robert W Kennard (1970). « Ridge regression : Biased estimation for nonorthogonal problems ». In : *Technometrics* 12.1, p. 55–67 (cf. p. 38, 40, 42).

Huang, Ke et Selin Aviyente (2007). « Sparse representation for signal classification ». In : *Advances in Neural Information Processing Systems 19 :Proceedings of the NIPS 2006 Conference*. MIT Press, p. 609–616 (cf. p. 86).

Ioannidis, John PA (2005). « Why most published research findings are false ». In : *PLoS medicine* 2.8, e124 (cf. p. 81).

Keshava, Nirmal et John F Mustard (2002). « Spectral unmixing ». In : *IEEE signal processing magazine* 19.1, p. 44–57 (cf. p. 23).

Kohavi, Ron et al. (1995). « A study of cross-validation and bootstrap for accuracy estimation and model selection ». In : *Ijcai*. T. 14. 2. Stanford, CA, p. 1137–1145 (cf. p. 50).

Komsta (2011). « Comparison of Several Methods of Chromatographic Baseline Removal with a New Approach Based on Quantile Regression ». en. In : *Chromatographia* 73.7-8, p. 721–731 (cf. p. 135).

Lee, Daniel D et H Sebastian Seung (1999). « Learning the parts of objects by non-negative matrix factorization ». In : *Nature* 401.6755, p. 788–791 (cf. p. 24).

Liddle, Andrew (2015). *An introduction to modern cosmology*. John Wiley & Sons (cf. p. 10).

Lin, Qiu-Hua, Jingyu Liu, Yong-Rui Zheng, Hualou Liang et Vince D Calhoun (2010). « Semiblind spatial ICA of fMRI using spatial constraints ». In : *Human brain mapping* 31.7, p. 1076–1088 (cf. p. 24).

Loncan, Laetitia, Luis B de Almeida, José M Bioucas-Dias, Xavier Briottet, Jocelyn Chanussot, Nicolas Dobigeon, Sophie Fabre, Wenzhi Liao, Giorgio A Licciardi, Miguel Simoes et al. (2015). « Hyperspectral pansharpening : a review ». In : *IEEE Geoscience and remote sensing magazine* 3.3, p. 27–46 (cf. p. 25).

Mallat, S. et Z. Zhang (1993). « Matching pursuits with time-frequency dictionaries ». In : *IEEE Trans. Signal Processing* 41.12, p. 3397–3415 (cf. p. 86).

Manolakis, Dimitris, Ronald Lockwood, Thomas Cooley et John Jacobson (2009). « Is there a best hyperspectral detection algorithm ? » In : *SPIE Defense, Security, and Sensing*. International Society for Optics et Photonics, p. 733402–733402 (cf. p. 78).

Manolakis, Dimitris G, Gary A Shaw et Nirmal Keshava (2000). « Comparative analysis of hyperspectral adaptive matched filter detectors ». In : *AeroSense 2000*. International Society for Optics et Photonics, p. 2–17 (cf. p. 78).

Meillier, Céline (2015). « Détection de sources quasi-ponctuelles dans des champs de données massifs ». Thèse de doct. Université Grenoble Alpes (cf. p. 79, 87).

Meillier, Céline, Florent Chatelain, Olivier Michel et Hacheme Ayasso. « Nonparametric Bayesian extraction of object configurations in massive data ». In : *IEEE Transactions on Signal Processing* 63.8, p. 1911–1924 (cf. p. 87).

— (2015). « Error control for the detection of rare and weak signatures in massive data ». In : *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, p. 1974–1978 (cf. p. 81, 95).

Meillier, Céline, Florent Chatelain, Olivier Michel, Roland Bacon, Laure Piqueras, Raphael Bacher et Hacheme Ayasso (2016). « SELFI : an object-based, Bayesian method for faint emission line source detection in MUSE deep field data cubes ». In : *A& A* 588, A140 (cf. p. xx, 9).

Meillier, Céline, Raphael Bacher, Florent Chatelain et Olivier Michel (2017). « Méthode de sigma-clipping par point fixe pour l'estimation de la distribution sous H0 dans le cadre de tests multiples ». In : *Colloque du Groupe de recherche et d'étude de traitement du signal et des images (GRETSI) 2017* (cf. p. xx, 119).

Moffat, AFJ (1969). « A theoretical investigation of focal stellar images in the photographic emulsion and application to photographic photometry ». In : *Astronomy and Astrophysics* 3, p. 455 (cf. p. 6, 29).

Nascimento, José MP et Jose MB Dias (2005a). « Does independent component analysis play a role in unmixing hyperspectral data ? » In : *IEEE Transactions on Geoscience and Remote Sensing* 43.1, p. 175–187 (cf. p. 24).

Nascimento, José MP et José MB Dias (2005b). « Vertex component analysis : A fast algorithm to unmix hyperspectral data ». In : *IEEE transactions on Geoscience and Remote Sensing* 43.4, p. 898–910 (cf. p. 24).

Ng, Andrew Y, Michael I Jordan, Yair Weiss et al. (2002). « On spectral clustering : Analysis and an algorithm ». In : *Advances in neural information processing systems* 2, p. 849–856 (cf. p. 146).

Ning, Xiaoran, Ivan W. Selesnick et Laurent Duval (2014). « Chromatogram baseline estimation and denoising using sparsity (BEADS) ». en. In : *Chemometrics and Intelligent Laboratory Systems* 139, p. 156–167 (cf. p. 135).

Paris, Sylvia, David Mary et André Ferrari (2013). « Detection tests using sparse models, with application to hyperspectral data ». In : *Signal Processing, IEEE Transactions on* 61.6, p. 1481–1494 (cf. p. 9, 79, 87, 97).

Pony, Olivier, Xavier Descombes et Josiane Zerubia (2000). « Classification d'images satellitaires hyperspectrales en zone rurale et périurbaine ». In : (cf. p. 148).

Poskitt, DS (1987). « Precision, complexity and Bayesian model determination ». In : *Journal of the Royal Statistical Society. Series B (Methodological)*, p. 199–208 (cf. p. 45).

Rafelski, Marc, Harry I Teplitz, Jonathan P Gardner, Dan Coe, Nicholas A Bond, Anton M Koekemoer, Norman Grogin, Peter Kurczynski, Elizabeth J McGrath, Matthew Bourque et al. (2015). « UVUDF : Ultraviolet Through Near-infrared Catalog and Photometric Redshifts of Galaxies in the Hubble Ultra Deep Field ». In : *The Astronomical Journal* 150.1, p. 31 (cf. p. 28, 30, 62).

Ramirez, Carlos, Vladik Kreinovich et Miguel Argaez (2013). « Why l1 Is a Good Approximation to l0 : A Geometric Explanation ». In : (cf. p. 44).

Reed, Irving S et Xiaoli Yu (1990). « Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution ». In : *Acoustics, Speech and Signal Processing, IEEE Transactions on* 38.10, p. 1760–1770 (cf. p. 78).

Rousseeuw, Peter J et Christophe Croux (1993). « Alternatives to the median absolute deviation ». In : *Journal of the American Statistical association* 88.424, p. 1273–1283 (cf. p. 46).

Rousseeuw, Peter J. et Annick M. Leroy (1987). *Robust regression and outlier detection*. New York : Wiley (cf. p. 135–137).

Rousseeuw, Peter J. et Katrien Van Driessen (2006). « Computing LTS regression for large data sets ». In : *Data mining and knowledge discovery* 12.1, p. 29–45 (cf. p. 137).

Scharf, Louis L et L Tood McWhorter (1996). « Adaptive matched subspace detectors and adaptive coherence estimators ». In : *Signals, Systems and Computers, 1996. Conference Record of the Thirtieth Asilomar Conference on*. IEEE, p. 1114–1117 (cf. p. 78).

Schowengerdt, Robert A (2006). *Remote sensing : models and methods for image processing*. Academic press (cf. p. 87).

Schwarz, Gideon et al. (1978). « Estimating the dimension of a model ». In : *The annals of statistics* 6.2, p. 461–464 (cf. p. 38, 44).

Serra, P, R Jurek et L Flöer (2012). « Using negative detections to estimate source-finder reliability ». In : *Publications of the Astronomical Society of Australia* 29.3, p. 296–300 (cf. p. 106).

Sherman, Jack et Winifred J Morrison (1949). « Adjustment of an inverse matrix corresponding to changes in the elements of a given column or a given row of the original matrix ». In : *Annals of Mathematical Statistics*. T. 20. 4, p. 621–621 (cf. p. 51).

Slepian, David (1962). « The One-Sided Barrier Problem for Gaussian Noise ». In : *Bell System Technical Journal* 41.2, p. 463–501 (cf. p. 134).

Solomon, J et B Rock (1985). « Imaging spectrometry for earth remote sensing ». In : *Science* 228.4704, p. 1147–1152 (cf. p. 78).

Soto, Kurt T., Simon J. Lilly, Roland Bacon, Johan Richard et Simon Conseil (2016). « ZAP – enhanced PCA sky subtraction for integral field spectroscopy ». In : *Monthly Notices of the Royal Astronomical Society* 458.3, p. 3210–3220 (cf. p. 4).

Storey, John D et al. (2003). « The positive false discovery rate : a Bayesian interpretation and the q-value ». In : *The Annals of Statistics* 31.6, p. 2013–2035 (cf. p. 82).

Storey, John D, Jonathan E Taylor et David Siegmund (2004). « Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates : a unified approach ». In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 66.1, p. 187–205 (cf. p. 94, 95).

Theobald, CM (1974). « Generalizations of mean square error applied to ridge regression ». In : *Journal of the Royal Statistical Society. Series B (Methodological)*, p. 103–106 (cf. p. 40).

Tibshirani, Robert (1996). « Regression shrinkage and selection via the lasso ». In : *Journal of the Royal Statistical Society. Series B (Methodological)*, p. 267–288 (cf. p. 38, 42).

Tikhonov, Andreï Nikolaevich, Vasiliï Iakovlevich Arsenin et Fritz John (1977). *Solutions of ill-posed problems*. T. 14. Winston Washington, DC (cf. p. 40).

Trujillo, I, JAL Aguerri, J Cepa et CM Gutiérrez (2001). « The effects of seeing on Sérsic profiles–II. The Moffat PSF ». In : *Monthly Notices of the Royal Astronomical Society* 328.3, p. 977–985 (cf. p. 7).

Vaart, Aad W Van der (1998). *Asymptotic statistics*. T. 3. Cambridge university press (cf. p. 132).

Villa, Alberto, Jocelyn Chanussot, Christian Jutten, Jon Atli Benediktsson et Saïd Moussaoui (2009). « On the use of ICA for hyperspectral image analysis ». In : *Geoscience and Remote Sensing Symposium, 2009 IEEE International, IGARSS 2009*. T. 4. IEEE, p. IV–97 (cf. p. 24).

Villeneuve, Emma (2012). « Déconvolution de données hyperspectrales pour l'instrument MUSE du VLT ». Thèse de doct. Université de Toulouse, Université Toulouse III-Paul Sabatier (cf. p. 6).

Villeneuve, Emma, Hervé Carfantan et Denis Serre (2011). « PSF estimation of hyperspectral data acquisition system for ground-based astrophysical observations ». In : *2011 3rd Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS)*. IEEE, p. 1–4 (cf. p. 6, 7, 29).

Wang, Xiang et Ian Davidson (2010). « Flexible constrained spectral clustering ». In : *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, p. 563–572 (cf. p. 148).

Wei, Qi, José Bioucas-Dias, Nicolas Dobigeon et Jean-Yves Tourneret (2015). « Hyperspectral and multispectral image fusion based on a sparse representation ». In : *IEEE Transactions on Geoscience and Remote Sensing* 53.7, p. 3658–3668 (cf. p. 25).

Weilbacher, Peter M, Ole Streicher, Tanya Urrutia, Aurélien Jarno, Arlette Pécontal-Rousset, Roland Bacon et Petra Bohm (2012). « Design and capabilities of the MUSE data reduction software and pipeline ». In : *SPIE Astronomical Telescopes+ Instrumentation*. International Society for Optics et Photonics, 84510B–84510B (cf. p. 3).

Winter, Michael E (1999). « N-FINDR : an algorithm for fast autonomous spectral end-member determination in hyperspectral data ». In : *SPIE's International Symposium on Optical Science, Engineering, and Instrumentation*. International Society for Optics et Photonics, p. 266–275 (cf. p. 24).

Xu, Qianqian, Ming Yan et Yuan Yao (2014). « Fast Adaptive Least Trimmed Squares for Robust Evaluation of Quality of Experience ». In : *arXiv preprint arXiv :1407.7636* (cf. p. 135, 136).

Yokoya, Naoto, Takehisa Yairi et Akira Iwasaki (2012). « Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion ». In : *IEEE Transactions on Geoscience and Remote Sensing* 50.2, p. 528–537 (cf. p. 25).

Yuan, Ming et Yi Lin (2006). « Model selection and estimation in regression with grouped variables ». In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 68.1, p. 49–67 (cf. p. 44).

Zhang, Zhi-Min, Shan Chen et Yi-Zeng Liang (2010). « Baseline correction using adaptive iteratively reweighted penalized least squares ». en. In : *The Analyst* 135.5, p. 1138 (cf. p. 135).

Résumé — Ces travaux se placent dans le contexte de l'étude des champs profonds hyperspectraux produits par l'instrument d'observation céleste MUSE. Ces données permettent de sonder l'Univers lointain et d'étudier les propriétés physiques et chimiques des premières structures galactiques et extra-galactiques. La première problématique abordée dans cette thèse est l'attribution d'une signature spectrale pour chaque source galactique. MUSE étant un instrument au sol, la turbulence atmosphérique dégrade fortement le pouvoir de résolution spatiale de l'instrument, ce qui génère des situations de mélange spectral pour un grand nombre de sources. Pour lever cette limitation, des approches de fusion de données, s'appuyant sur les données complémentaires du télescope spatial Hubble et d'un modèle de mélange linéaire, sont proposées, permettant la séparation spectrale des sources du champ. Le second objectif de cette thèse est la détection du Circum-Galactic Medium (CGM). Le CGM, milieu gazeux s'étendant autour de certaines galaxies, se caractérise par une signature spatialement diffuse et de faible intensité spectrale. Une méthode de détection de cette signature par test d'hypothèse est développée, basée sur une stratégie de max-test sur un dictionnaire et un apprentissage des statistiques de test sur les données. Cette méthode est ensuite étendue pour prendre en compte la structure spatiale des sources et ainsi améliorer la puissance de détection tout en conservant un contrôle global des erreurs. Les codes développés sont intégrés dans la bibliothèque logicielle du consortium MUSE afin d'être utilisables par l'ensemble de la communauté. De plus, si ces travaux sont particulièrement adaptés aux données MUSE, ils peuvent être étendus à d'autres applications dans les domaines de la séparation de sources et de la détection de sources étendues.

Mots clés : démélange spectral, fusion de données, inférence à grande échelle, hyperspectral, tests multiples, contrôle global d'erreurs.

Abstract — This work takes place in the context of the study of hyperspectral deep fields produced by the European 3D spectrograph MUSE. These fields allow to explore the young remote Universe and to study the physical and chemical properties of the first galactical and extra-galactical structures. The first part of the thesis deals with the estimation of a spectral signature for each galaxy. As MUSE is a terrestrial instrument, the atmospheric turbulences strongly degrades the spatial resolution power of the instrument thus generating spectral mixing of multiple sources. To remove this issue, data fusion approaches, based on a linear mixing model and complementary data from the Hubble Space Telescope are proposed, allowing the spectral separation of the sources. The second goal of this thesis is to detect the Circum-Galactic Medium (CGM). This CGM, which is formed of clouds of gas surrounding some galaxies, is characterized by a spatially extended faint spectral signature. To detect this kind of signal, an hypothesis testing approach is proposed, based on a max-test strategy on a dictionary. The test statistics is learned on the data. This method is then extended to better take into account the spatial structure of the targets, thus improving the detection power, while still ensuring global error control. All these developments are integrated in the software library of the MUSE consortium in order to be used by the astrophysical community. Moreover, these works can easily be extended beyond MUSE data to other application fields that need faint extended source detection and source separation methods.

Keywords : spectral unmixing, data fusion, large-scale inference, hyperspectral, multiple testing, global error control.

Gipsa-lab, 11 rue des Mathématiques,
38400 Saint-Martin d'Hères