Projet 80|Prime OrionStat incluant le financement d'une thèse 2020-2023.

Début souhaité au 1er octobre 2020.

Sujet : **Méthodes statistiques pour l'inversion de modèle et distribution spatiale des propriétés physico-chimiques du nuage moléculaire Orion B**

Encadrants / contacts :
Pierre Chainais, professeur à Centrale Lille / CRIStAL UMR 9189 ; **pierre.chainais@centralelille.fr**
Franck Le Petit, astronome, Observatoire de Paris - PSL, LERMA UMR 8112 ; Franck.LePetit@obspm.fr

Résumé :

Les nouveaux détecteurs millimétriques en astrophysique fournissent des masses de données qu'il n'est plus possible d'analyser avec les méthodes classiques. De plus, les modèles numériques utilisés pour interpréter ces observations produisent eux-mêmes de grands volumes de données, hétérogènes et en grande dimension. L'interprétation des observations de régions de formation stellaire avec les modèles de référence ne peut se faire qu'en inventant de nouvelles méthodes de traitement statistique du signal et de machine learning. Ce projet vise à surmonter deux verrous : 1) pouvoir résoudre des problèmes inverses sur des millions de pixels et parfois peu contraints, 2) réussir à dé-mélanger dans les observations interstellaires et extragalactiques les composantes émettrices pour estimer les paramètres physiques individuels. Il s'appuie sur un Large Program du TGIR IRAM, Orion-B. Les méthodes seront publiques via l'un des services nationaux d'observations de l'INSU.

Mots-clés : images hyper-spectrales, machine learning, problèmes inverses, formation stellaire, masses de données, incertitudes, intervalles de confiance

Sujet détaillé :

La formation des étoiles est un processus fondamental qui gouverne l'évolution des galaxies aux échelles de temps cosmiques. Tandis que ces dernières années ont apporté des avancées majeures pour la compréhension des mécanismes de formation des étoiles à l'échelle des galaxies et à l'intérieur des nuages moléculaires géants (GMC, giant molecular clouds), il reste néanmoins à mettre en cohérence les résultats obtenus dans la Voie Lactée avec ceux obtenus dans d'autres galaxies. La difficulté principale vient de l'extraordinaire gamme d'échelles spatiales à considérer. Les études à l'intérieur de notre galaxie résolvent les nuages interstellaires individuellement, ce qui permet la mise en relation entre propriétés physiques intrinsèques (densité volumique du gaz, nombre de Mach…) et l'activité de formation des étoiles à l'intérieur d'un nuage (e.g., Motte et al. 2018, Lee et al. 2016). Les observations extragalactiques n'accèdent pas à la structure des nuages individuels mais demeurent le meilleur moyen d'étudier le lien entre les propriétés des gaz, le taux de formation des étoiles, les paramètres globaux de la galaxie et l'environnement de la galaxie (e.g., Saintonge et al. 2017, Usero et al. 2015).

Les raies d'émission moléculaires et leurs ratios sont couramment utilisés pour déduire les propriétés des régions intra- et extra-galactiques de formation des étoiles. La nouvelle génération de capteurs sub-millimétriques large bande comme la caméra EMIR sur l'antenne 30m de l'IRAM permet d'accéder aux détails d'un grand nombre de raies spectrales. Ce projet a pour objectif de développer des méthodes avancées de traitement du signal et d'apprentissage statistique (machine learning) adossées à des modèles physico-chimiques avancés pour les appliquer aux observations issues d'Orion-B (Outstanding Radio-Imaging of OrioN-B : http://iram.fr/~pety/ORION-B/) afin d'en extraire des informations physiques aussi fines et précises que possible. Le télescope est situé dans la Sierra Nevada en Espagne où une équipe scientifique internationale menée par Jérôme Pety, astronome de

l'Observatoire de Paris, en poste à l'IRAM, a obtenu les observations radio les plus complètes du nuage moléculaire géant (GMC) Orion B (Gratier et al. 2017, Leroy et al. 2017, Pety et al. 2017). Ces observations vont permettre de comprendre comment les parties internes les plus denses et les plus froides du nuage donnent naissance aux étoiles. Connu pour abriter les nébuleuses de la Tête de Cheval et de la Flamme, Orion B est un gigantesque réservoir de matière interstellaire, de gaz et de poussières, qui contient environ 70 000 fois la masse du soleil. Les endroits où les futures étoiles peuvent naître, dits cœurs denses, brillent uniquement aux longueurs d'onde radio millimétriques : ils sont invisibles aux télescopes optiques. L'instrumentation récente à l'IRAM 30m a permis d'obtenir des images 100 fois plus grandes qu'avant, et cela à de très nombreuses longueurs d'onde millimétriques en même temps, et avec une excellente résolution spatiale grâce à un balayage fin du champ angulaire couvert.

Les données du projet ORION-B éclairent une des questions clés de l'astrophysique moderne, à savoir pourquoi les nuages moléculaires forment-ils si peu d'étoiles ? Alors que les nuages devraient s'effondrer sous leur propre poids pour se transformer entièrement en cœur dense puis en étoile, seuls quelques pourcents du nuage se transforment en réalité en étoile. Le projet ORION-B délivre environ 160 000 images, ou encore 1h50 de film à 24 images par secondes. Il ne fait aucun doute que ce type d'observations radio va se généraliser dans un futur proche. Ainsi, les approches statistiques pionnières que nous proposerons dans le cadre de ce projet apporteront les outils et l'expérience nécessaires pour faire parler des jeux de données de plus en plus riches et de grande dimension.

Dans ce contexte où aucune vérité terrain n'est disponible, il s'agira d'imaginer des nouvelles stratégies efficaces pour estimer les paramètres physiques pertinents à partir d'images hyper-spectrales de très grande dimension (820 000 pixels, 240 000 canaux spectraux/pixel) tout en étant capable d'assortir ces estimations d'intervalles de confiance assortis de garanties théoriques. Les données sont des spectres acquis dans une zone de 5 degrés carrés (environ 60 années lumières de large) autour du nuage moléculaire Orion B avec une résolution angulaire de 26'' (0.1 année lumière). ORION-B récolte les données autour d'au moins 30 raies spectrales moléculaires dans la gamme 72-116 GHz avec une résolution de 0,6 $km.s^{-1}$ environ (rappelons ici qu'on utilise couramment une traduction des fréquences en vitesse). Cette gamme spectrale inclut des traceurs moléculaires galactiques et extragalactiques usuels tels que CO, HCO, HCN, CS. Le cube de données hyper-spectrales issu de ces mesures est unique au vu de la masse d'informations qu'il contient, laissant espérer pour la première fois une caractérisation de la structure physique, chimique et dynamique d'un nuage moléculaire géant (GMC) en lien avec l'activité de formation des étoiles.

Exploiter à plein la richesse d'un tel jeu de données nécessite d'imaginer de nouvelles méthodes en science des données, adaptées à un contexte extrême en termes de dimensions. L'ambition de ce projet est de reformuler les questions sur la formation des étoiles en termes de problèmes de traitement du signal et d'apprentissage statistique qui prennent en compte les limitations du système d'observation telles que la projection de structures tridimensionnelles sur le plan du ciel. Les méthodes qui cherchent à répondre à ce type de questions font l'objet de recherches actives actuellement (Pereyra 2017, Repetti et al. 2019, Vono, Dobigeon, Chainais 2019). L'objectif est d'extraire les informations cachées dans des volumes de données qui dépassent les capacités humaines d'exploration visuelle, aux limites mêmes de l'état de l'art en science des données. Il s'agit aussi de fournir de nouvelles références pour les simulations numériques de l'évolution des nuages moléculaires géants. La très grande gamme de valeurs des conditions physiques et le rapport signal-à-bruit de nos données combiné à leur très grand volume constituent un véritable défi. Nous disposons pour cela de modèles numériques qui encodent les propriétés physico-chimiques. Dans le projet Orion B ce travail est réalisé à l'Observatoire de Paris par F. Le Petit et E. Bron, qui développent le code *Meudon PDR* (Le Petit et al. 2006, Bron et al. 2014). Ce code, l'une des références largement utilisé par la communauté, simule de façon couplée le transfert de rayonnement, la chimie et les processus thermiques dans une tranche de gaz interstellaire afin d'estimer des intensités de raies d'émission théoriques en fonction des paramètres physiques du nuage.

Méthodes statistiques pour l'inversion de modèle et distribution spatiale des propriétés physico-chimiques du nuage moléculaire Orion B

L'extraction d'information utile à partir des données grand champ hyper-spectrales issues d'ORION-B, complexes, massives et bruitées, implique la relation à des modèles complets pour inférer la distribution spatiale des paramètres physiques (pression, température, rayonnement UV lointain,…) et chimiques (abondances, fraction ionisée, déplétions, …) du nuage. Les méthodes utilisées pour cette étape d'inférence basée sur les modèles doivent tenir compte des dégénérescences potentielles associées à un problème inverse mal posé qui pourraient amener à sur-interpréter les données. La variabilité du rapport signal-à-bruit rend la détermination d'estimateurs précis difficile. Disposer d'incertitudes quantifiées de façon fiable est essentiel à une interprétation physique robuste. Nous tirerons partie pour cela de la grande taille des données du projet ORION-B qui se transforme alors en avantage pour étudier et valider nos approches, en particulier pour assortir nos prédictions d'intervalles de confiance. Nous nous intéresserons surtout à des approches qui permettent de produire des intervalles de confiance ou de crédibilité accompagnés de garanties théoriques puisque nous ne disposons ici d'aucun oracle, aucune vérité terrain. Certains de nos travaux vont déjà dans cette direction (Vono, Dobigeon, Chainais 2019 ; Vono, Dobigeon, Chainais 2020)

Approches bayésiennes pour l'inférence des cartes de distribution des paramètres physico-chimiques avec intervalles de crédibilités garantis.

Le niveau de bruit étant parfois bas et le problème inverse étant mal posé, nous devons faire appel à des a priori pour définir l'espace dans lequel nous cherchons nos estimateurs. Les modèles astrophysiques dont nous disposons, tels que le code PDR de Meudon, sont souvent très coûteux en temps de calcul. Par conséquent, nous utiliserons des grilles de modèles pré-calculés. Ensuite nous utiliserons des méthodes de Monte Carlo à chaînes de Markov, en particulier de type ABC (Approximate Bayesian Computation) pour inclure la connaissance a priori de la physique du système dans notre méthode d'inférence. Les méthodes ABC sont conçues pour traiter les situations où la vraisemblance ne peut être explicitement calculée lors de la recherche de la solution au problème inverse (Wilkinson 2013) : la grille de modèles pré-calculée est alors utilisée comme une boîte noire pour explorer l'espace des paramètres et évaluer leur probabilité a posteriori. L'une des difficultés sera d'ailleurs l'exploitation optimale d'une interpolation efficace de cette grille. L'algorithme final devra 1) fournir les solutions les plus probables décrivant les cartes de paramètres physico-chimiques, 2) fournir des intervalles de confiance aux estimateurs, si possible assortis de garanties théoriques. L'approche proposée devra aussi permettre de résoudre les régions à faible SNR. Ces recherches s'inscriront naturellement dans le champ de nos travaux récents sur ces questions (Vono et al. 2019).

Pour les cartes du milieu interstellaire (ISM ou InterStellar Medium) nous pouvons de plus utiliser des a priori assez forts : les conditions physiques au sein d'un GMC sont spatialement régulières et ne devraient pas faire apparaître de variations erratiques d'un pixel à l'autre. Une régularisation spatiale bien choisie sera nécessaire. Ce type d'approche est commune en traitement d'image où l'on utilise habituellement des régularisations de type gradient ou Laplacien, ou encore TV (total variation) éventuellement généralisée. Nous explorerons aussi le potentiel de régularisations impliquant la parcimonie dans un espace de représentation telle que les ondelettes, adaptées au contrôle du poids de textures multi-échelles isotropes dans les images. Nous envisageons aussi de considérer les approches d'apprentissage profond bayésien. Nos approches seront testées aussi bien sur des cartes synthétiques que sur des observations réelles pour identifier les a priori les plus adaptés, tout en tenant compte de la complexité numérique qui devra passer à l'échelle de la masse de données à exploiter. Nous devrons aussi étudier l'influence d'une éventuelle mauvaise spécification du modèle puisqu'il aura servi à la fois d'a priori dans la formulation du problème inverse et de référence pour les tests de validation : l'erreur d'approximation due à une mauvaise spécification du modèle se traduit

nécessairement par une contribution à l'erreur d'estimation. Les cartes obtenues pour des paramètres tels que la densité, densité de colonne, illumination UV, fraction ionisée ou la température permettront de mieux comprendre et analyser les conditions physiques à l'intérieur du nuage Orion B.

Démélange des contributions moléculaires.

Un défi général posé par l'interprétation de ce type d'analyse naît du mélange des contributions de différents types de milieux interstellaires aux raies d'émission observées. Ce problème est critique pour l'étude des régions de formations stellaires dans d'autres galaxies, où en raison du manque de résolution spatiale, de nombreux milieux sont mélangés dans le lobe du télescope. Ce mélange superpose (de façon linéaire ou non-linéaire) les contributions de plusieurs milieux le long de la même ligne de visée dans les observations. L'interprétation de ces mesures intégrées spatialement le long de ligne de visée à partir d'un seul jeu de paramètres physiques peut mener à des conclusions erronées, ce que ne s'interdisent pas de faire certaines études qui négligent cet aspect. Nous comptons nous attaquer à ce verrou en exploitant la richesse et la haute résolution des données issues de la campagne ORION-B.

Il s'agira d'identifier les multiples composantes émettrices sur une ligne de visée, comme des PDRs, des chocs, du gaz sombre ou diffus. L'objectif est de déterminer les conditions physiques dans chacune des composantes. Sans démélange, on obtient un résultat erroné - souvent supposé correspondre à un résultat « moyen » dans la littérature. Le problème auquel on se confronte est hautement dégénéré, et deux types d'*a priori* seront proposés. D'abord, la connaissance des composantes physiques elles-mêmes et des spécificités de leur émission (grâce aux modèles astrophysiques) sera très importante. Les données Orion-B fournissent un exemple de GMC spatialement résolu qui servira de test. Elles serviront aussi à définir des *a priori* sur la forme des distributions de conditions physiques démélangées. Cette séparation des composantes répond à des questions astrophysiques cruciales : en schématisant, si dans une galaxie, on conclut que ce sont des chocs qui dominent, cela signifie que le milieu peut être turbulent et que s'y sont produites de nombreuses explosions de supernovae. Tandis qu'une forte contribution des PDRs fournit une information quantitative sur le taux de formation stellaire. Les méthodes développées seront aussi appliquées aux observations du programme PHANGS (Large Program MUSE et ALMA).

Rôle du doctorant.

Le doctorant recruté devra surtout apporter une expertise en traitement du signal et apprentissage statistique. Il devra aussi avoir l'esprit ouvert et une certaine curiosité pour l'astrophysique. Son rôle sera crucial puisqu'il devra d'une part appréhender le jeu de données ORION-B dans toutes sa complexité et sa richesse, en comprendre les enjeux en termes de physique du nuage moléculaire, et proposer des méthodes adaptées aux objectifs décrits plus haut. Il sera localisé au laboratoire CRIStAL à Lille et se rendra régulièrement à l'Observatoire de Paris, la communication est aisée entre ces deux villes.

Dans un premier temps, il devra poursuivre deux études bibliographiques en parallèle. La première consistera en un état de l'art sur les problèmes inverses en grande dimension, en particulier dans le cadre bayésien puisque nous cherchons des estimateurs assortis d'intervalles de confiance ; il devra aussi s'intéresser aux garanties théoriques associées. Une deuxième étude bibliographique portera sur la partie astrophysique, notamment en lien avec les observations, la physico-chimie et les modèles numériques des nuages moléculaires. Il est probable que l'acclimatation à un sujet interdisciplinaire tel que celui-ci nécessite un investissement important en début de thèse. Notre équipe dispose des compétences nécessaires et sera attentive à l'accompagnement du doctorant dans ce travail qui établira les fondations du projet de thèse. Le doctorant bénéficiera de l'expertise internationale disponible au sein du consortium ORION-B.

Ensuite, en fait dès que sa culture le permettra, le doctorant pourra se familiariser avec les algorithmes de l'état de l'art et leur adaptation à notre problématique. Il devra progressivement monter en puissance pour devenir force de proposition de nouvelles méthodes qui prennent en compte tous les attendus mais aussi toutes les contraintes de ce projet. Nous tâcherons d'identifier une première approche de reconstruction des cartes de paramètres physico-chimiques qui permette de mettre en valeur l'investissement du début de thèse d'ici la fin de la première année. Ce travail se concrétisera par l'écriture d'un premier article de journal. Il participera aussi à la production des grilles de modèles pour l'interprétation des observations. Participer à cette seconde tâche lui donnera une expertise sur la complexité de ces modèles et leur degré d'incertitude inhérent qu'il est parfois difficile de quantifier, mais qu'il faut avoir en tête lorsqu'on interprète les observations.

En deuxième année, le doctorant devrait pouvoir aborder des méthodes plus avancées, impliquant un bon compromis entre complexité numérique, précision des estimateurs, et intervalles de confiance garantis. En parallèle, il pourra s'attaquer au problème du démélange des lignes de visée qui semble poser plus de difficultés techniques.

En troisième année, l'expertise interdisciplinaire acquise par le doctorant devrait lui permettre de proposer une méthode avancée combinant les tâches de démélange et d'identification des cartes de paramètres physico-chimiques. Idéalement, un troisième article de revue devrait être soumis avant d'engager la rédaction du manuscrit.

Ce travail s'appuie sur le Large Program Orion B de l'IRAM, l'un des TGIR du ministère. L'un des objectifs de ce Large Program est de développer de nouvelles méthodes statistiques pour exploiter les masses de données que fournissent désormais les radio-télescopes. Un objectif important de cette thèse sera la diffusion des méthodes développées auprès de la communauté astrophysique. Nous mettrons à disposition l'ensemble des codes sources produits, dans l'esprit d'une science ouverte et reproductible.

Profil recherché : Le candidat devra avoir des bases solides en traitement statistique du signal et des images et/ou en machine learning, avec un attrait pour l'astrophysique. Un M2 dans l'une de ces disciplines est demandé. Des notions en résolution de problèmes inverses serait un plus. Des compétences de programmation (Python notamment) seront nécessaires.

Equipe d'encadrement.

Pierre Chainais est un expert en traitement statistique du signal et apprentissage statistique. Physicien de formation, il a l'habitude de travailler avec des astronomes et astrophysiciens depuis de nombreuses années. Il a précédemment travaillé avec l'Observatoire Royal de Belgique (SoHo, étude sur le Soleil calme), et collabore actuellement avec le consortium ORION-B ainsi qu'avec le Laboratoire AstroParticule et Cosmologie (APC, Paris) au sujet de l'analyse des ondes gravitationnelles. **Pierre-Antoine Thouvenin** est un jeune Maître de Conférences spécialisé dans les méthodes d'optimisation et de Monte Carlo pour la résolution de problèmes inverses. Il a fait un postdoc à Edimbourg dans une équipe spécialisée en imagerie astronomique (Institute of Sensors, Signals & Systems, Heriot Watt Univ.). **Franck Le Petit** est un Astronome confirmé, très impliqué dans le projet ORION-B. Il connaît parfaitement ces données ainsi que les codes de simulation physico-chimiques Meudon PDR. **Emeric Bron** est un astronome adjoint spécialiste de ces modèles physico-chimique ; il a récemment intégré le LERMA et fait partie du consortium ORION-B. Nous avons tous profité du **PEPS AstroInfo 2018-2019** pour initier cette collaboration et apprendre à nous parler entre disciplines différentes. Cette première exploration nous as permis de consolider les fondations de ce projet 80Prime.

Références

Gratier, P. et al. 2017, Astronomy & Astrophysics, 599, 100

Lee, E. et al. 2016, Astrophysical Journal, 833, 229L

Motte, F. et al. 2018, Nature Astronomy, 2, 478

Pereyra, M., 2017, SIAM Journal on Imaging Sciences, vol. 10, no. 1, 285–302.

Pety J. et al. 2017, Astronomy & Astrophysics, 599, 98

Repetti, A., Pereyra, M., Y. Wiaux, 2020, SIAM Journal on Imaging Sciences, vol. 12, no. 1, pp. 87-118, 2019.

Saintonge, A., et al. 2017, Astrophysical Journal Supplemental Series, 233, 22

Usero, A. et al 2015, Astrophysical Journal, 150, 115

Vono, M., Dobigeon N., Chainais P., 2019, IEEE Transactions on Signal Processing, 66, 17, 4541.

Vono, M., Dobigeon N., Chainais P., 2020, preprint .

Wilkinson, R. (2013), Statistical applications in genetics and molecular biology, 12, 1–13