UNIVERSITÉ PARIS-SUD

ÉCOLE DOCTORALE ONDES ET MATIERE

T H È S E

préparée en vue d'obtenir le titre de

Docteur en Sciences

de l'Université Paris Sud XI

par

Othman BENOMAR

Sismologie Stellaire : Méthodes statistiques appliquées aux étoiles de type solaire

Thèse dirigée par Frédéric BAUDIN

soutenue le 21 Octobre 2010

Jury :

Rapporteurs :	Yveline LEBRETON	-	GEPI, Observatoire de Paris
	John LEIBACHER	-	National Solar Observatory (USA)
Directeur :	Frédéric BAUDIN	-	Institut d'Astrophysique Spatiale
Président :	Guillaume PINEAU DES FORETS	-	Institut d'Astrophysique Spatiale
Examinateurs :	Thierry APPOURCHAUX	-	Institut d'Astrophysique Spatiale
	Yvonne ELSWORTH	-	University of Birmingham (UK)
Invité :	Michael BAZOT	-	Universidade do Porto (Portugal)

Remerciements

Ce travail de thèse n'aurait pu s'accomplir sans le soutien de mon tuteur Frédéric Baudin. Je tiens à le remercier vivement pour sa grande patience, ses conseils et son écoute. Je remercie aussi Thierry Appourchaux, guide suprême des statistiques en astérosismologie sans qui je n'aurais pu aboutir. Ces conseils avisés m'ont été d'une aide inestimable.

C'est à Marie-Jo Goupil, Juao Pedro Marques, Réza Samadi ,Yveline Lebreton et Marc antoine Dupret que je dois aussi ce travail. Sans eux, je me serais perdu dans les méandres de la modélisation. Je les remercie vivement pour leur patience et leur écoute.

Je n'oublie pas non plus l'accueil très chaleureux du groupe HIROS avec qui j'ai passé des moments inoubliables. Je remercie en particulier Bill Chaplin et Yvonne Elsworth pour le temps qu'ils ont pu me consacrer entre deux sandwiches et un café. Leurs conseils et leur soutien m'ont été d'une grande aide.

Un autre merci à toute l'équipe CoRoT : techniciens, ingénieurs et scientifiques. Parmis ceux-ci, j'ai une pensé toute particulière envers Annie Baglin et Eric Michel qui ont porté CoRoT à maturation et sans qui mon travail de thèse n'aurait pu être effectué.

Et bien sûr, je remercie mes collègues de bureau à l'IAS, Clément, Nicolas, Céline, Romain et Claudia qui, ont su être patient devant mes divagations et râles journaliers. Je n'oublierais pas non plus les échanges fructueux et chaleureux que nous avons pu avoir.

Merci aux membres du service informatique pour les heures qu'ils ont pu passer à tenter de maintenir en vie le réseau informatique. Je remercie aussi les ordinateurs, pour les longues heures de souffrances que je leur ai fait subir. Sans eux, je n'aurais pu accomplir tout ce travail.

Je remercie également mes amis, Guillaume Boudharam et David Charalampous pour les échanges instructifs que nous avons pu avoir sur les divers domaines des sciences.

Enfin, à tous mes amis et proches pour les moments de rire et de détente que nous avons pu partager ! longue vie et prospérité.

Science et science-fiction ne sont jamais très loin l'un de l'autre. Je finirais donc ces remerciements par un petit clin d'oeil à ce qui m'amena à l'astrophysique, il y a fort, fort longtemps :

Space : The final frontier
These are the voyages of the Starship, Enterprise
Its 5 year mission
To explore strange new worlds
To seek out new life and new civilizations
To boldly go where no man has gone before [1].

1. Tiré de la saga Star Trek.

Table des matières

1 Introduction **1**

1.1 Un peu d'histoire... 1

1.1.1 CoRoT, de ses prédécesseurs et successeurs 3

1.2 Eléments de physique stellaire 6

1.2.1 Généralités sur la structure interne et l'évolution stellaire 6

1.2.2 Equations et structure interne 11

1.2.3 Pulsations stellaires et bestiaire d'ondes 14

1.2.4 Propriétés des oscillations : approximation de Cowling 16

1.2.5 Modes p, modes g et modes mixtes 17

1.2.6 Influence de la rotation et visibilité des modes 23

1.2.7 Sur les mécanismes d'excitation des modes 24

1.2.8 Relation empirique sur les étoiles de type solaire 27

2 Sur les méthodes d'analyse et de traitement de signal classiquement usitées en astérosismologie **31**

2.1 Vraisemblance et approche MLE 31

2.1.1 Généralités 31

2.1.2 Fonction de vraisemblance d'un spectre de puissance 33

2.1.3 Comparaison de modèles avec une approche MLE 34

2.2 Approche Bayésienne simple : Approche MAP 35

2.2.1 Probabilité *a priori* et probabilité *a posteriori* 35

2.2.2 Application de l'approche Bayésienne au cas stellaire 36

2.2.3 Comparaison de modèles avec une approche MAP 37

2.3 Le calcul d'erreur avec des algorithmes de type MLE ou MAP 38

3 L'échantillonnage par Markov Chain Monte-Carlo **41**

3.1 Introduction sur les méthodes d'échantillonnage 41

3.1.1 Echantillonnage par réjection 42

3.1.2 Qu'est ce qu'un processus Markovien ? 43

3.2 L'Echantillonage de Monte-Carlo par Chaine de Markov 44

3.2.1 Définitions et propriétés : des chaines de Markov à l'algorithme de Metropolis-Hasting 44

3.2.2 Quel(s) critère(s) pour assurer une exploration optimale de l'espace des paramètres ? 47

3.2.3 L'Adaptive Markov Chain Monte-Carlo : un moyen de contrôle automatique du taux d'acceptation/réjection 54

3.2.4 Eviter le confinement dans des maxima locaux du critère statistique 61

3.2.5 Comparaison de modèles avec un algorithme MCMC tempéré 64

3.2.6 Résumé sur la méthode d'échantillonnage 68

4 Méthodes, simulations et applications des MCMC sur des étoiles de CoRoT 71
4.1 Quel méthodologie adopter ? 71
4.1.1 Une stratégie globale d'ajustement du spectre de puissance 71
4.1.2 Une approche bayésienne pour les pulsateurs de type solaire 73
4.2 Simulations 76
4.2.1 Fonction de densité de probabilité et choix de critères statistiques 76
4.2.2 Spectres stellaire simulés 86
4.3 Applications aux étoiles 91
4.3.1 Un cas stellaire intensivement étudié : HD49933 91
4.3.2 Un autre cas difficile : HD181420 100

5 Extraction automatique et systématique de paramètres stellaires 113
5.1 Problématique 113
5.2 Extraction d'information générale sur les modes 114
5.2.1 Méthodologie 114
5.2.2 Les connaissances *a priori* utilisées 115
5.3 Extraction des modes 115
5.3.1 Comment automatiser l'ajustement des modes ? 116
5.3.2 Contraintes de continuité des fréquences 117
5.3.3 Contrainte de continuité des hauteurs et largeurs 120
5.3.4 Le bruit lors de l'ajustement des modes 123
5.3.5 Remarque sur l'angle d'inclinaison et le splitting rotationnel 123
5.3.6 Résumé de la méthode 124
5.4 Test sur les données solaires : mise en évidence du cycle solaire 125
5.5 Performance de l'algorithme sur un échantillon d'étoiles simulées 130

6 De l'analyse du signal à son interpretation physique 143
6.1 CESAM, LOSC et minimisation 144
6.2 HD49933 146
6.2.1 Physique 147
6.2.2 Méthode 147
6.2.3 Discussion des résultats 148
6.3 HD181420 153

7 Conclusion et perspectives 161
7.1 Quelques mots sur les différences entre fréquentistes et bayésiens 161
7.2 L'extraction de modes de pulsation 162
7.3 Des modes mixtes à foison 163
7.4 Interprétation : une critique de la méthode 165
7.5 Perspectives à moyen et long terme 166

A ANNEXE : bibliothèque de fonctions *a priori* 175

B ANNEXE : tables de résultats (paramètres *a posteriori*) pour HD49933 181

C ANNEXE : tables de résultats (paramètres *a posteriori*) pour HD181420 187

Chapitre 1

Introduction

Sommaire

1.1 Un peu d'histoire... 1
1.1.1 CoRoT, de ses prédécesseurs et successeurs 3
1.2 Eléments de physique stellaire 6
1.2.1 Généralités sur la structure interne et l'évolution stellaire 6
1.2.2 Equations et structure interne 11
1.2.3 Pulsations stellaires et bestiaire d'ondes 14
1.2.4 Propriétés des oscillations : approximation de Cowling 16
1.2.5 Modes p, modes g et modes mixtes 17
1.2.6 Influence de la rotation et visibilité des modes 23
1.2.7 Sur les mécanismes d'excitation des modes 24
1.2.8 Relation empirique sur les étoiles de type solaire 27

1.1 Un peu d'histoire...

D'aussi loin que l'archéologie nous permet de remonter, l'espèce humaine a observé le ciel et les astres qui le composent tentant de percer le mystère de leur création. Stonehenge dont l'axe du monument est aligné avec le Soleil au solstice d'été ou encore les multiples dieux dédiés au Soleil d'anciennes Égypte en sont de bonnes illustrations. Mais bien souvent, ce fut au travers d'un voile mystico-religieux que les hommes ont étudié les astres.

Bien que certains philosophes et lettrés purent formuler les lois de la physique avec des moyens rudimentaires [1], ce n'est qu'avec le développement des premiers instruments capables de voir au delà de ce que l'oeil peut percevoir que l'astronomie et les sciences qui lui sont rattachées ont pu effacer peu à peu le caractère sacré du ciel. Ainsi, la lunette de Galilée (1564 - 1642) inventé en 1609, constitue la révolution scientifique sans laquelle nous ne pourrions prétendre faire d'astérosismologie ni même d'astronomie.

La variabilité de certaines étoiles a été mise en évidence depuis presque aussi longtemps que l'invention du téléscope. Mira (une étoile géante rouge), observé par David Fabricius (1564 - 1617) à la fin du 16ème siècle puis par Johannes Phocylides Holwarda (1618 - 1651) au début de 17ème fut ainsi la première étoile pour laquelle les variations lumineuses au cours du temps ont été établies de manière univoque. Mais il fallut attendre le développement de la photographie

1. Citons par exemple les lois de Kepler, découlant d'observations effectuées par Tycho Brahé à l'aide d'astrolabes et de quadrants.

à la fin du 19ème siècle pour que des mesures quantitatives de luminosités de qualité puissent être effectuées. A cette époque, des questions telles que les mécanismes physiques en jeu dans les étoiles étaient âprement discutées. Il fallut attendre les développements théoriques de la physique atomique au début du 20ème pour enfin comprendre d'où elles tirent leur énergie.

Parallèlement, le nombre d'étoiles variables n'a cessé de s'accroitre et c'est Arthur Eddington (1882 - 1944) qui le premier proposa au courant des années 1930 une interprétation en terme de pulsation stellaire. A la même époque, c'est grâce à une classe bien particulière de pulsateurs, les céphéides (nommés en référence à l'étoile prototype δ *cephei*), que l'une des avancées majeures en astrophysique a été possible. Ces étoiles pulsent de manière périodique avec des variations de magnitude pouvant aller jusqu'à plusieurs unités. Leavitt & Pickering (1912) ont établi une dépendance directe entre la période de pulsation et la luminosité de ces étoiles. Or cette dépendance présuppose que chacune de ces étoiles a une masse et une température identique. En tant que telle, elles peuvent servir d'étalon cosmique et ont ainsi permis de sonder l'espace à des échelles qui n'avaient jamais été possible d'atteindre jusque là. Par exemple, les nuages de Magellan que l'on pensait autrefois situés à l'intérieur de notre galaxie se sont révélés bien plus lointain. C'est en fait une galaxie satellite de la nôtre. Par delà cette considération sur l'univers local, Hubble (1889 - 1953) a pu grâce à ces étalons stellaires mettre en évidence l'expansion de l'univers, contredisant ainsi la vision d'Einstein (1879 - 1955) d'un univers statique, obligeant ce dernier à abandonner la constante cosmologique introduite dans la relativité générale de manière ad-hoc. Einstein qualifiera cette constante cosmologique comme étant la plus grande erreur de sa vie.

Les pulsations stellaires ont donc permis une révision radicale de la vision du monde en cosmologie. Mais revenons à des considérations plus terriennes : le Soleil. Attribuable à Leighton et al. (1962) et Evans & Michard (1962) les mesures d'oscillations en surface du Soleil sont plutôt tardives. L'interprétation de ces oscillations de surface en terme d'oscillation globale a été effectuée de manière indépendante par Ulrich (1970) et Leibacher & Stein (1971) puis confirmée expérimentalement par Deubner et al. (1975). Cette découverte a permis pour la première fois de sonder l'intérieur même du Soleil. Tout l'art de l'héliosismologie consiste alors à mesurer ces oscillations et d'en déduire (processus d'*inférence*, aussi appelé *inversion*) sur la base de considérations d'ordre physique la structure interne du Soleil. En effet, la vitesse moyenne du son dans la région pénétrée par un mode peut être déduite des fréquences du-dit mode. Ainsi l'évaluation des propriétés d'un ensemble de modes traversant différentes régions du Soleil (plus particulièrement grâce à la fréquence) permet par *inversion héliosismique* de reconstruire les profils de variations de la vitesse du son, de la température et de la pression au sein de l'étoile.

Les conséquences de l'héliosismologie ne sont pas limitées à la compréhension de la physique stellaire. Cette science a permis par exemple de mettre indirectement en évidence l'oscillation des neutrinos. Durant les années 1980, la structure interne obtenue par inversion a montré que le coeur du Soleil, dont la température est contrôlée par le taux de réaction nucléaire, était bien plus chaud que ce que laissait supposer le flux de neutrino mesuré de manière indépendante (déficit de neutrino sur Terre en provenance du Soleil). Plusieurs explications ont été produites pour expliquer ce désaccord manifeste, dont l'hypothèse d'oscillation des neutrinos. Cette hypothèse considère que les 3 *saveurs* existantes de neutrinos[2] peuvent s'interchanger par interaction (très faible) avec la matière. Certains ont aussi suggéré des solutions plus exotiques tel que l'ajout de nouvelles particules de masse non nulle, interagissant peu avec la matière : les WIMPs (*Weakly Interacting Massive Particules*). Au cours des ans, la qualité des données et des modèles hélio-

2. Les neutrinos ont été découvert en 1956 et les neutrinos cosmiques mesurés qu'en 1987 par Koshiba suite à l'explosion de la supernova SN1987A. Il fallu attendre 1998 pour que l'oscillation des neutrinos soit réellement établit.

sismiques s'améliorant, le problème de déficit de neutrinos persistait. Ce ne fut qu'en 1998 que, pour la première fois les oscillations de neutrinos ont put être mises en évidence expérimentalement (expérience Super-Kamiokande au Japon). Le déficit de neutrinos solaires est maintenant imputé à un changement de saveur des neutrinos au cours de leur trajet vers la terre.

Ces dernières années, le nombre d'étoiles variables connues s'est accru significativement grâce à des moyens techniques toujours plus performants et sensibles. Il est maintenant certains que toutes (ou presque) les étoiles pulsent et c'est le rôle de l'astérosismologie de débusquer, identifier et interpréter les oscillations stellaires. Il ne fait aucun doute que les progrès futurs en astérosismologie permettront de bien mieux comprendre la structure des intérieurs stellaires sur l'ensemble du diagramme HR.

De nos jours l'astérosismologie se développe sur tous les fronts. Du sol avec par exemple SONG (Stellar Observations Network Group) qui est le pendant stellaire de réseaux d'observations solaires comme BISON (créé dans les années 1980, BISON fut l'un des premiers du genre). SONG est en cours d'élaboration et vise à étendre un réseau de stations automatisées de par le globe. Ce type de réseau permettrait d'observer une large gamme d'étoile par vélocimétrie avec précision, sur de longues échelles de temps et de manière continue. En orbite, avec par exemple MOST (Walker et al. 2003) ou CoRoT (Baglin et al. 2006), satellite du CNES dédié à la recherche d'exoplanète et à l'astérosismologie. Au delà, avec Kepler (Borucki et al. 2007) qui est le pendant de CoRoT à la NASA et dans le futur probablement PLATO (Catala et al. 2010).

1.1.1 CoRoT, de ses prédécesseurs et successeurs

Notre travail de thèse ayant essentiellement porté sur l'étude des étoiles CoRoT (Convection, Rotation and Planetary Transits), une brève présentation du satellite s'impose. CoRoT est la première mission spatiale dont les objectifs avoués sont l'astéroseismologie et l'exoplanétologie. Contrairement à MOST (Microvariability and Oscillations of Stars), qui est un microsatellite (60kg) canadien dédié à l'astéroseismologie, CoRoT est une mission de plus grande envergure, d'une masse de 670 kg.

Le satellite a été placé en orbite le 27 décembre 2006 a 896 km d'altitude sur une orbite polaire pour une durée nominale de mission de 2.5 ans. Dans la voie astérosismologie, une dizaine d'objets de magnitude 6 a 9.5 sont étudiés simultanément grâce à 2 des 4 cameras CCD [3]. Les temps d'observations doivent être les plus long possibles pour obtenir une résolution suffisante dans le spectre de puissance. C'est pourquoi des cycles d'observations de 20 jours à 150 jours ont été définis assurant une résolution spectrale allant de 0.57 μHz à 0.08 μHz.

Détaillons quelques caractéristiques de l'instrument :

Le télescope. Il est constitué d'un système de miroirs de pupille d'entrée de 27 cm. Un baffle cylindrique permet d'éviter toute lumière parasite, notamment la lumière diffusée provenant de la Terre. Le bruit instrumental est donc théoriquement limité à quelques photons/pixel/seconde. Un obturateur permet de protéger le télescope avant et pendant le lancement. Le télescope pointe une région de d'environ 12° de rayon appelée les yeux des CoRoT, située à l'intersection du plan galactique et de l'équateur céleste.

Les cameras CCD. Elles sont essentiellement constituées d'une matrice de 2048×4095 pixels, et conçues pour fonctionner dans le visible (de 370 nm à 950 nm). Leur température est maintenue à $-40°$ C avec des variations n'excédant pas 0.015°, pour limiter le bruit de fond aléatoire.

3. Suite à un problème de DPU, survenue en Février 2009, le satellite ne dispose plus que de la moitié de ces CCD.

Un objectif de 1200 mm de focale permet aussi de corriger les aberrations chromatiques et géométriques du télescope. Ces capteurs servent à la fois les objectifs Exoplanètes et Astérosismologie. Compte tenu des exigences de la partie exoplanètes, une partie des CCD reçoit un faisceau de lumière après avoir traversé un prisme. Observer les différentes composantes du spectre permet en effet de distinguer les variations lumineuses induites par l'activité stellaire de celles induites par un transit planétaire (l'activité stellaire étant dépendante de la longueur d'onde, contrairement au transit).

La baie d'équipement. Elle contient toute l'instrumentation et l'équipement nécessaires au contrôle de température, de télémétrie ou encore les unités de stockage des données. Les fluctuations de température pouvant changer quelque peu les capacités de traitement des circuits électroniques, celle ci est précisément régulée.

Le logiciel de bord. Il est indispensable pour maintenir tous les systèmes de CoRoT en bon état de marche. Certaines tâches de prétraitement des données lui sont aussi imputées.

Alors que l'on n'espérait observer sismiquement qu'une centaine d'étoiles de magnitude comprise entre 6 et 9.5, la mission a à ce jour, mesurée les pulsations de milliers d'étoiles. En effet, malgré de plus faibles magnitudes ciblées (11 à 16) la voie dédiée aux exoplanètes s'est aussi révélée exploitable pour sonder les pulsations sur près d'un millier d'étoile comme par exemple, les géantes rouges (De Ridder et al. 2009; Hekker et al. 2009) et les étoiles chaudes de type B (Degroote et al. 2009).

Plusieurs étoiles de type solaire ont été observées et analysées. On citera par exemple, HD49933 (Appourchaux et al. 2008), HD181420 (Barban et al. 2009), HD49385 (Deheuvels et al. 2010), HD175726 (Mosser et al. 2009b) et HD181906 (García et al. 2009). Certaines de ces étoiles ont fait l'objet d'une ré-analyse dans le cadre de notre thèse (cf. Chap.4) et suivi de publications (Benomar 2008; Benomar et al. 2009a,b). En dehors du volet astérosismologie, CoRoT observe également les variations lumineuses induites par des transits planétaires. C'est ainsi plus de 50000 étoiles qui ont été scrutées pour y détecter d'éventuels compagnons planétaires d'une taille pouvant descendre jusqu'à quelques rayons terrestres. Au total c'est plus d'une douzaine de planètes qui a été mis au jour, dont la première planète de type super-terre historiquement découverte : CoRoT 7b (Léger et al. 2009).

Au final, cette mission a été un véritable succès produisant des données de qualité et en nombre, permettant pour ce qui concerne l'astérosismologie, des analyses statistiques des comportements stellaires sans précédent sur tout le diagramme HR. Elle a ainsi ouvert la voie à une nouvelle ère concernant la compréhension des processus physiques en jeu au sein des étoiles.

CoRoT est toujours en service puisque son fonctionnement a été prolongé de 3 ans au delà de sa durée initialement prévue. Il a été rejoint par Kepler le 7 Mars 2009. Cette mission est encore plus ambitieuse que CoRoT puisqu'elle observera durant 3 années (durée nominale qui sera très probablement prolongée) un champ fixe du ciel d'environ 12 degrés carré contenant plus de 400 0000 étoiles (dont 100 000 observées durant 3 ans). Les premières analyses montrent un comportement incroyablement riche et inattendu pour bon nombre de pulsateurs. C'est aussi un défi pour la communauté astérosismique car l'échantillon d'étoiles est monstrueux et pose des défis quant à l'analyse d'une telle quantité de données. Malheureusement, pour des questions de politique de publication nous ne pourrons présenter pleinement les résultats que nous avons déjà obtenus grâce à Kepler. Cependant c'est sans conteste que nous pouvons affirmer que l'apport de cette mission sera majeure dans les prochaines années.

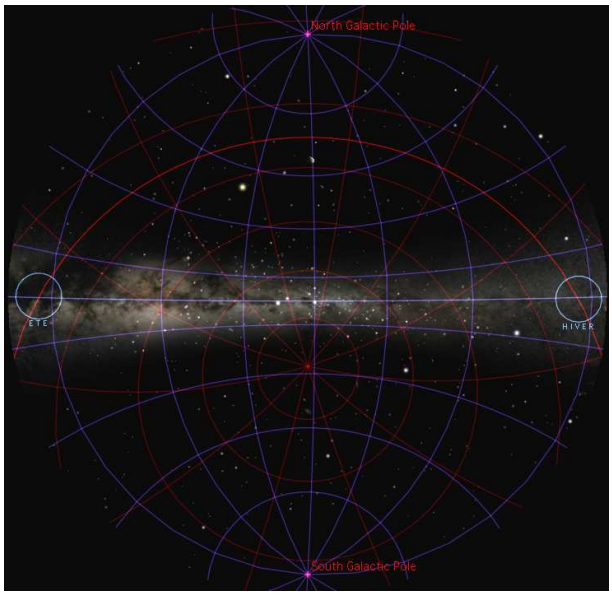


FIGURE 1.1 – Champs de vision de CoRoT (petits cercles). En été le satellite observe une zone proche de la constellation de l'Aigle (à gauche) tandis qu'en hiver il observe une zone voisine de la constellation d'Orion (à droite).

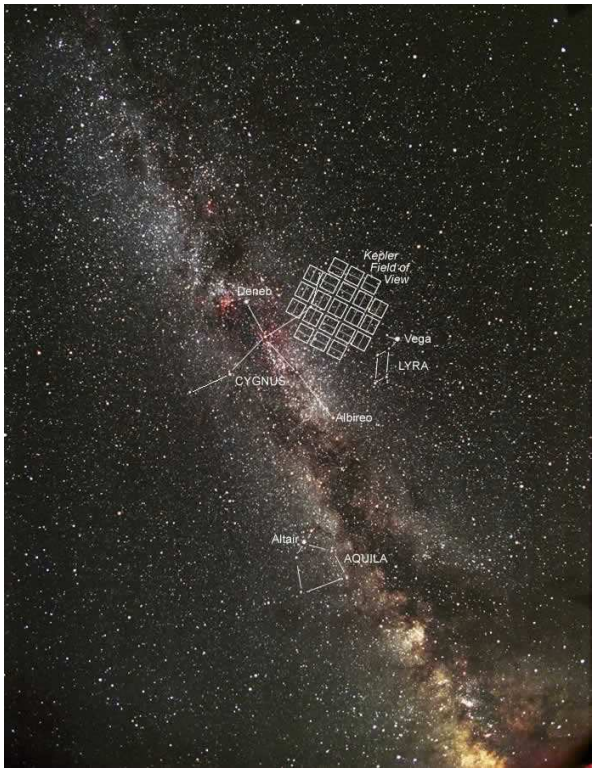


FIGURE 1.2 – Champ de vision de Kepler. Chacun des carrés représentent le champ couvert par l'un des 42 CCD.

1.2 Eléments de physique stellaire

Cette section vise à présenter une vue d'ensemble de ce que sont les étoiles dans toute leur diversité. Commençons par donner une définition de ce qu'est une étoile.

1.2.1 Généralités sur la structure interne et l'évolution stellaire

En première approximation une étoile peut être vue comme une boule de gaz ionisé maintenu à l'équilibre thermodynamique (à une échelle de temps humaine) et hydrostatique grâce à l'action et la réaction des forces de gravité et de pressions, ces dernières étant issues essentiellement des réactions de fusion au coeur. Cette définition très générale inclut une large variété d'étoiles allant de quelques fractions de masse solaire (masse très légèrement supérieures à $\simeq 0.07M_{\odot}$, limite supérieure des naines brunes) à une centaine de masse solaire (étoiles chaudes de type spectrale O). Tout autre corps entre soit dans la catégorie des planètes ($\lesssim 13M_{jov}$), des naines brunes[4] ($13M_{jov} \lesssim M \lesssim 0.07M_{\odot}$) ou des étoiles mortes (naines blanches, étoiles à neutron et éventuelles trou-noir).

Tous ces corps sont susceptibles de pulser et font l'objet de recherche. Notons que les planètes ne font pas exception puisque récemment des campagnes de mesures au sol ont permis de mettre en évidence des oscillations globales de Jupiter (Gaulme et al. 2010).

Diagramme de Hertzsprung-Russell

Le diagramme de Hertzsprung-Russell (diagramme HR, Fig.1.3) aussi connu sous le nom de diagramme couleur-magnitude, représente la luminosité des étoiles en fonction de leur température. Des étoiles de masse et de séquence évolutive (âge) différentes dessinent un motif dans le diagramme HR permettant une visualisation aisée des différents types/classes d'étoiles. Plusieurs sections peuvent être clairement distinguées. La plus caractéristique d'entre elle est la barre diagonale partant des étoiles les plus froides et moins brillantes, aux étoiles les plus chaudes et brillantes. Toutes les étoiles de cette structure font partie de la "séquence principale" phase durant laquelle elles brulent leur hydrogène au coeur. Les moins massives sont situées en bas à droite tandis que les plus massives sont situées en haut à gauche (plus chaudes et lumineuses). Dans la partie droite supérieure du diagramme se trouvent les étoiles ayant quitté la séquence principale, plus agées et en phase géante. La partie froide et faiblement lumineuse du diagramme (en bas à gauche) regroupe les astres "éteints" que sont les naines blanches.

Au cours de leur évolution, les étoiles explorent différentes régions du diagramme HR. Initialement, un nuage de matière froide se condense sous l'effet de la force de gravitation et s'échauffe. De là nait au bout de quelques millions d'années une *proto-étoile* au centre du nuage et des corps résiduels. Lors de l'effondrement, le nuage de matière est en rotation. Du fait de la conservation du moment cinétique, une proto-étoile tourne relativement rapidement. Au cours de sa vie, la rotation ralentit (perte de moment cinétique par friction et interaction gravitationnelle avec les corps voisins). De l'échauffement dû à la contraction émerge un rayonnement qui s'accentuera au cours du temps jusqu'à ce que celle-ci amorce ces réactions de fusion (initialement *via* des réactions proton-proton -PP-) et acquiert le statut d'étoile. Cette phase est appelée pré-séquence principale (*pre main sequence*) ou PMS et se termine par la ZAMS (*zero age main sequence*), terme associé à l'instant à partir duquel l'étoile rejoint la séquence principale. Elle y restera la plus grande partie de sa vie, sa luminosité et sa température s'accroissant très lentement (traversé

4. La frontière entre naine brune et planète ou, entre naine brune et étoile n'est pas clairement déterminée. Ces corps célestes sont supposés entretenir leur luminosité (classes spectrales M, L, T, Y) de part la chaleur interne lors de leur création, mais ne sont pas le siège de réaction nucléaire majeures sur de longues échelles de temps.

du diagramme HR perpendiculairement à la diagonale formant la séquence principale : la température de surface décroit tandis que la luminosité augmente). Au bout d'une durée qui peut varier de quelques millions d'années (étoiles les plus massives) à plusieurs dizaines de milliards d'années (étoiles les moins massives), il n'y a plus d'hydrogène au coeur (en terme évolutif, cet instant est appelé TAMS pour *terminal age main sequence*). Les étoiles les plus massives commencent à bruler de l'hélium, puis de plus en plus rapidement tous les éléments lourds qu'il lui est permis de bruler dans une série de cycle de fusion. Les étoiles les plus massives (*i.e.* $M > 8M_\odot$) fusionnent des éléments jusqu'à former les éléments du pic du fer. Les moins massives (*i.e.* $M < 0.5M_\odot$) s'arrêtent à la fusion de l'hélium. Dans les étoiles de masse intermédiaire comme le Soleil, les conditions de pression et de température ne permettent pas de produire d'éléments plus lourd que le carbone. Tous les éléments plus lourds que le fer sont produits lors de l'explosion des étoiles ou en phase calme dans les enveloppes des étoiles supergéantes de la branche asymptotique (AGB). Toutes ces réactions se font essentiellement au travers de 3 types de réactions dont l'importance relative dépend de la phase évolutive et des conditions de pression et de température que l'étoile peut atteindre (donc de sa masse). On distingue ainsi les chaînes de réaction PP, la réaction 3α et le cycle CNO. De l'hydrogène est aussi brulé par strate dans les couches supérieures. Durant cette phase l'étoile quitte donc graduellement la séquence principale pour rejoindre le groupe des géantes. A l'arrêt de toutes les réactions nucléaires, l'équilibre pression/gravité est définitivement rompu. L'étoile s'effondre sur elle même, à la recherche d'un nouvelle équilibre. Celui ci est garanti par la pression de dégénérescence des électrons pour les naines blanches et par la pression de dégénérescence des neutrons pour les étoiles à neutron. Pour les trous noir, il n'existe pas de force pour contrecarrer la gravité ce qui aboutit à une singularité (toute la masse serait virtuellement concentrée en un point de l'espace).

Le Soleil et les étoiles de type solaire dans la séquence principale

Le Soleil en tant qu'étoile la mieux connue est une excellente référence pour comprendre les mécanismes et caractéristiques types au sein des étoiles de masse voisine dans la séquence principale. De plus, ce sont justement les étoiles de type solaire sur lesquelles cette thèse s'est focalisée. Il n'est donc pas inutile de décrire plus en détail le Soleil.

Le Soleil est un plasma ovoïde (aplatissement très faible aux pôles dû à la rotation de tel sorte que l'hypothèse de sphéricité du plasma est valide) dont environ 74% de la masse est constituée d'hydrogène et 24% d'hélium. Le résiduel (1.77%) est une *mixture* d'éléments chimiques plus lourds, souvent appelés *métaux*. Son rayon moyen est de $6,955.10^8$ m pour une masse de $1,99.10^{30}$ kg et une densité moyenne de $1,408.10^3$ kg/m^3, ce qui est a peine plus que l'eau. Soulignons qu'il est difficile de définir de manière non ambigüe la surface du Soleil car ce que nous appelons surface correspond à la zone à partir de laquelle l'atmosphère devient transparente au rayonnement. Mais cette donnée peut varier en fonction de la longueur d'onde d'observation.

Sa structure interne est constituée d'un coeur dans lequel les réactions de fusion entre noyaux d'hydrogène produisent de l'hélium. Le coeur s'étend sur environ 20 à 25% de rayon solaire pour une densité d'environ 150.10^3 kg/m^3 et une température avoisinant les 15 millions de Kelvin. Les zones les plus profondes du soleil ($R < 0.7R_\odot$) transportent leur energie de manière essentiellement radiative, comme illustré Fig.1.4. Ce n'est pas forcement le cas pour des étoiles légerement plus massives ($\simeq 1.1 - 1.3M_\odot$) : un coeur convectif peut apparaitre. L'absence de convection dans le couches profondes solaires provoque une migration par sedimentation des élements chimiques. Les abondances des différentes espèces chimiques se distribuent en fonction de la température, pression et densité et donc, du rayon de l'étoile. Dans la zone radiative solaire, les grandeurs physiques varient fortement. La température et la densité chutent respectivement

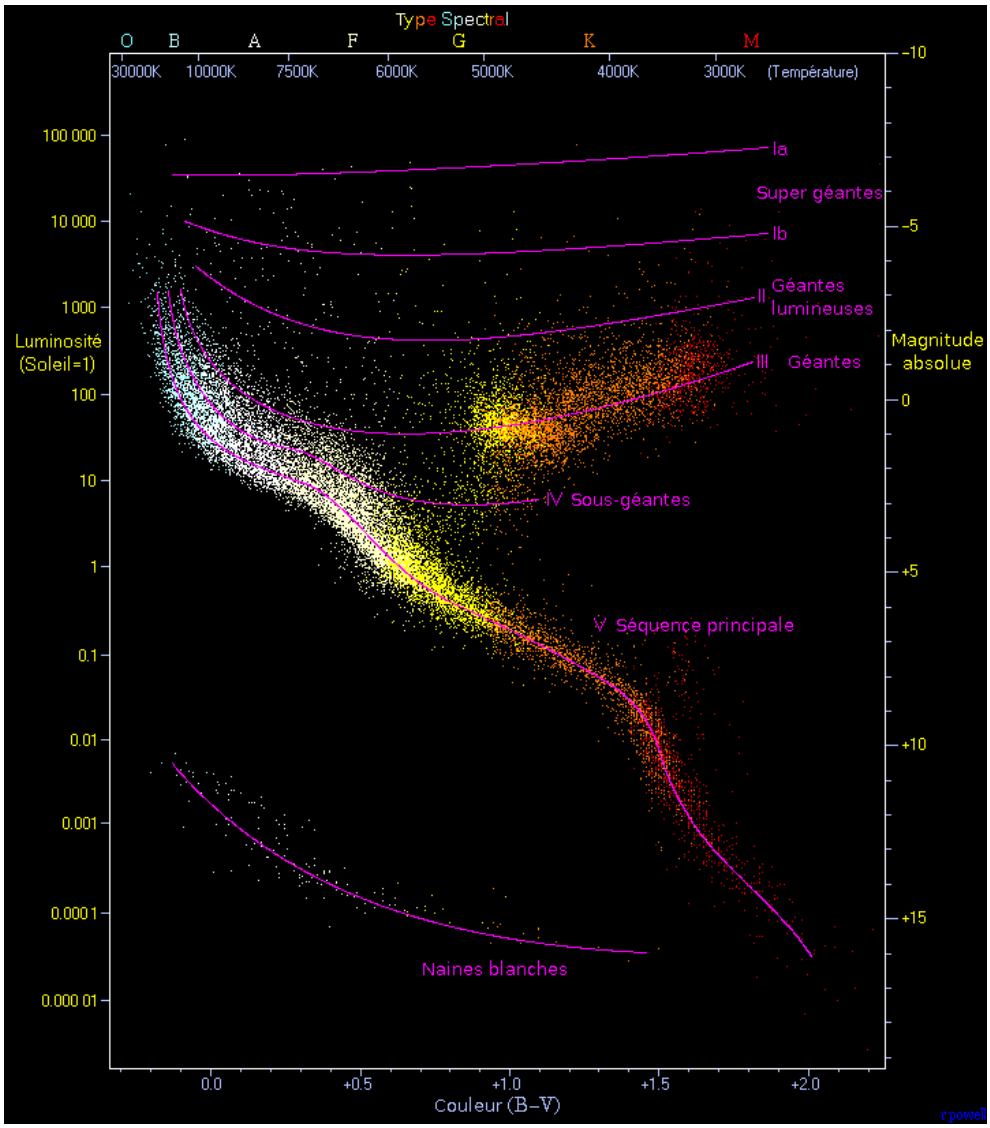


FIGURE 1.3 – Diagramme HR sur lequel sont représentées 22000 étoiles du catalogue Hipparcos et 1000 autres du catalogue Gliese. Les étoiles les plus chaudes et brillantes se situent en haut à gauche du diagramme tandis que les plus froides et moins lumineuses se trouvent en bas à droite. Le Soleil est situé sur la séquence principale pour une température approximative de 5777 K et une luminosité de 1. Ce diagramme a été initialement produit par Powell (2009).

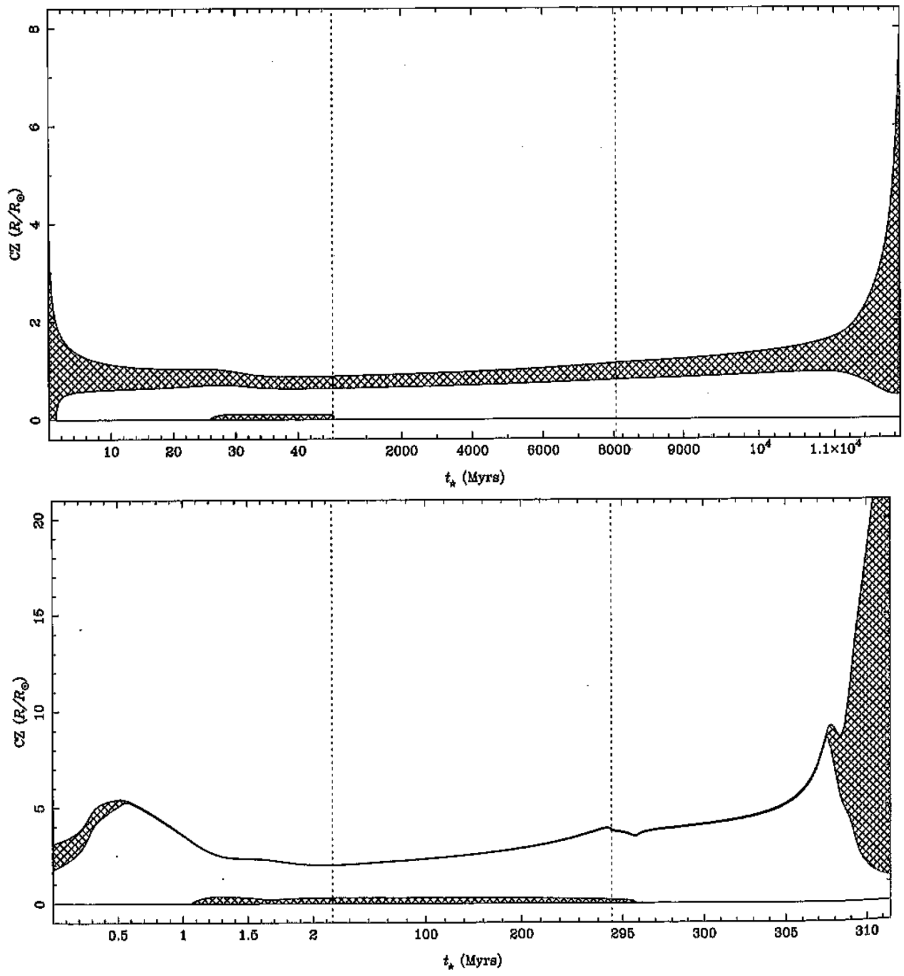


FIGURE 1.4 – Evolution temporelle des étendues convectives (zones grisées) pour une étoile de 1 masse solaire (haut) et pour une étoile de 3 masses solaires (bas), obtenue avec le code d'évolution stéllaire CESAM. La séquence principale est délimitée par les lignes pointillées verticales (respectivement la ZAMS et la TAMS). Illustration tirée de Marques et al. (2008).

jusqu'à 2 millions de Kelvin et 200 kg/m^3.

Les couches les plus externes du Soleil transportent essentiellement la chaleur grâce à des mouvements de convection. Contrairement à la zone radiative, les différentes espèces chimiques sont continuellement brassées dans la zone convective. Il en résulte une distribution d'espèce très homogène, indépendante de la profondeur.

Dans les zones les plus profondes du Soleil, on pense que la rotation est solide (indépendante de la latitude et du rayon). Ceci est bien établi pour les étendues les plus externes de la zone radiative. Par contre au niveau du coeur, les incertitudes des mesures héliosismiques sont grandes et cette supposition est sujette à discussion. Il est par contre clairement établi que les couches supérieures (dans la zone convective) du Soleil passent d'une zone de rotation différentielle marquée en surface, à une rotation solide à la base de la zone convective. La région de transition, soupçonnée d'être la source du champ magnétique solaire est appelée la *tachocline*.

En moyenne la surface du soleil est portée à 5777 Kelvin pour une densité de 2.10^{-4} kg/m^3.

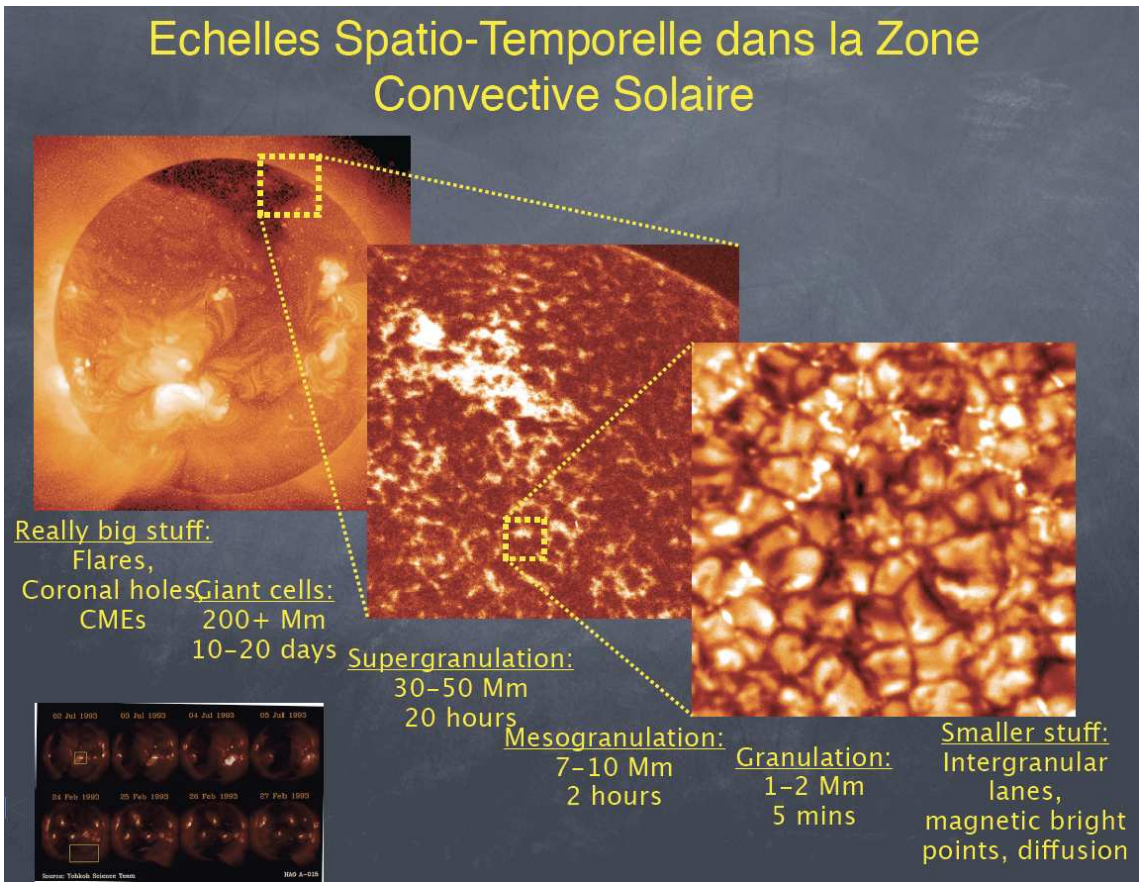


FIGURE 1.5 – Les différentes échelles convectives et structures éruptives en surface du Soleil. Image issue du cours de Tahar Amari.

Phénomène	τ (s)	σ (m/s)	p
Granulation	$\simeq 100$	$\simeq 0.7$	2
Meso-granulation	$\simeq 2000$	$\simeq 0.3$	2
Super-granulation	$\simeq 10^4$	$\simeq 1.9$	2
Régions actives	$\simeq 10^6$	$\simeq 3.0$	4

TABLE 1.1 – Ordre de grandeurs des paramètres de profils de Harvey pour les différentes structures à la surface du Soleil (Aigrain et al. 2004).

Le bruit de granulation et les régions actives de surface

A la surface du Soleil, des phénomènes convectifs sont visibles (Fig.1.5) sous forme de *granules*. Toutes les étoiles ayant une enveloppe supérieure convective ont ce type de structure de surface ($M \lesssim 2.5 M_\odot$). Un granule correspond à une zone de remontée de plasma chaud des profondeurs vers la surface. Autour de cette zone, le flux de matière est descendant, c'est à dire qu'un flux de matière froide pénètre dans les couches les plus profondes de la surface de l'étoile. Par effet de contraste, les zones de surface les plus chaudes apparaissent brillantes et les plus froides sombres, ce qui donne cette structure en damier de la Fig.1.5. Il est difficile de chiffrer la taille et la durée de vie des granules car il existe tout un spectre quasi continue de tailles de structures convectives allant de quelques centaines de kilomètres (granulation) à une cinquantaine de mégamètres (supergranulation). Les temps de vie caractéristiques associées vont de quelques minutes à plusieurs dizaines d'heures. Au fur et à mesures que l'on s'enfonce dans la surface de l'étoile, les structures granulaires deviennent de plus en plus étendues. Les phénomènes convectifs précédemment décrits se manifestent aussi dans les séries temporelles.

Une autre source de périodicité du signal photométrique solaire provient des régions actives. Les régions actives sont des zones de champ magnétique intense. Selon le scénario le plus probable, un champ intense émerge sous forme de boucles des profondeurs (par un mécanisme non clairement élucidé). En surface, la matière est entrainée et confinée dans les boucles (théorème du gel d'Alfven) : les mouvements de convection sont inhibés par le champ. En conséquence, les bases des boucles correspondent à des zones de plasma plus froid que le milieu environnant (jusqu'à 2000 Kelvin de moins) : ce sont les *taches solaires/stellaires*.

Dans le spectre de puissance, les différents phénomènes de surface sont assez bien reproduits par une semi-lorentzienne généralisée,

$$P(\nu) = \frac{4\sigma^2\tau}{1+(2\pi\nu\tau)^p} = \frac{A}{1+(B\nu)^p}, \tag{1.1}$$

avec σ l'intensité du processus, τ son temps de vie caractéristique et p un paramètre dépendant de la nature du phénomène et de son degré de cohérence spatiale et temporelle. Cette description empirique a été utilisée initialement par Harvey (1985) pour décrire les spectres de puissance solaire avec $p = 2$. En fait, dans le cas solaire on peut montrer qu'une description de l'ensemble du spectre de puissance nécessite au moins 3 profils de Harvey auxquels correspondent les différentes échelles caractéristiques : granulation, meso/super-granulation, région actives. On a résumé Table 1.1, les grandeurs caractéristiques des profils de Harvey généralisés, dans le cas du Soleil.

1.2.2 Equations et structure interne

La structure interne d'une étoile ne peut pas être observée par ondes électromagnétiques à cause de l'opacité du milieu. Cependant, en se basant sur des principes physiques, la structure

interne peut être déduite. Traitant le gaz constituant l'étoile comme un continuum, ces propriétés physiques locales (densité ρ, pression P, vitesse $\boldsymbol{v}$, ...) peuvent être définies par des fonctions du vecteur position $\boldsymbol{r}$ et du temps t. L'hydrodynamique, la thermodynamique et la loi de Poisson nous permettent d'établir 4 équations fondamentales (écrites en Eulérien, dans une base sphérique) décrivant la dynamique du système "Etoile" :

- la conservation de la masse,

$$\frac{\partial \rho}{\partial t} + \nabla.(\rho \boldsymbol{v}) = 0. \tag{1.2}$$

- la conservation du mouvement,

$$\rho \left(\frac{\partial}{\partial t} + \boldsymbol{v}.\nabla \right) \boldsymbol{v} = -\nabla P + \rho \nabla \Phi. \tag{1.3}$$

  Φ représente le potentiel gravitationnel, tel que $\boldsymbol{g} = -\nabla\Phi$.
- la loi de Poisson liant champ et matière. Appliqué à un champ de gravité on a

$$\Delta\Phi = 4\pi G\rho \tag{1.4}$$

  où G est la constante de gravitation.
- la conservation de l'énergie, déduite directement de la thermodynamique,

$$\rho T \left(\frac{\partial}{\partial t} + \boldsymbol{v}.\nabla \right) S = \rho\epsilon - \nabla.\boldsymbol{F}. \tag{1.5}$$

  T, S, ϵ, $\boldsymbol{F}$ désignent respectivement la température, l'entropie, le taux de production massique d'énergie et le flux d'énergie. Pour une étoile de type solaire, la source d'énergie est issue des réactions nucléaires et le flux dépend de la nature du transport de chaleur dominant : conduction, convection ou radiation.

Ces équations ne tiennent pas compte des forces électromagnétiques et des effets de la rotation (force de Coriolis et d'entrainement du fluide). En les incluant, on brise la symétrie sphérique ce qui rend bien plus complexe les calculs de modèles car grâce à la symétrie sphérique la structure interne de l'étoile ne dépend que du rayon (modèle unidimensionnel). Leur prise en compte se fait la plupart du temps *a posteriori*, sur la base d'un développement perturbatif. Sauf comportements extrêmes (effets relativistes, vitesse de rotation proche de vitesse de libération, ...) ces équations et suppositions permettent déjà d'obtenir une description de la dynamique stellaire satisfaisante dans le cas solaire et probablement pour une grande part des étoiles de type solaire. Le chap.1.2.6 traitera de l'effet de la rotation, celle ci étant la plupart du temps la principale perturbation de la structure interne d'équilibre d'une étoile.

On fera remarquer que le système d'équation est encore incomplet puisque nous avons 6 variables pour 4 équations. Deux relations de fermeture doivent être introduites.

La première porte sur la nature du flux thermique. Par exemple dans les zones radiatives, le flux convectif est nul et le milieu peut être supposé optiquement épais. L'absorption/ré-émission du rayonnement par le milieu n'est plus négligeable. De plus le transfert par conduction est très mauvais dans un plasma non dégénéré (c'est le cas dans les étoiles de la séquence principale). Dans ces conditions le flux radiatif total normalement issue de la concurrence entre la conduction, la radiation et la convection se fait exclusivement par radiation. Dans une zone radiative la loi de Fourier nous permet d'écrire,

$$F \equiv F_{RAD} = -\lambda_{RAD}\nabla T = -\frac{4ac}{3\kappa\rho}T^3\nabla T. \tag{1.6}$$

Ici, a, c et κ sont respectivement la constante de densité de radiation, la vitesse de la lumière et l'opacité. Bien qu'il soit très difficile d'établir un comportement moyen de l'opacité (de Rosseland), celle ci diminue en générale lorsque la température croit selon une loi de puissance,

$$\kappa_R = \kappa_0 \rho^\alpha T^\beta. \tag{1.7}$$

Dans les intérieurs stellaires, le milieu est complètement ionisé. Dans ce cas, cette loi est appelée loi de Kramers, pour laquelle $\alpha = 1$, $\beta = -7/2$, κ_0 dépend de la nature de l'élément chimique. Cette tendance est importante car elle amortit les effets de perturbations dans le flux d'énergie. Ainsi, si par exemple, la température d'une enveloppe de gaz de composition chimique donnée s'accroit, son opacité va décroitre et moins d'énergie sera déposée par radiation sur cette enveloppe, ce qui par réaction rabaisse la température. Il peut toutefois exister des espèces chimiques et des régimes d'ionisation pour lesquelles la dépendance température/opacité n'aboutit pas à cette auto-régulation du flux déposé, mais à son augmentation.

On peut trouver une description plus formelle de la notion d'opacité dans Cox (1984). En général, l'opacité est fournie par des tables que l'on interpole en fonction de la composition chimique (abondance en hydrogène X, en Hélium Y et en métaux Z), de la densité et de la température du plasma. Peu d'équipes sont capables de calculer ces tables de par la complexité des calculs en oeuvre. On citera par exemple les travaux de Rogers & Iglesias (1996) et de Badnell et al. (2005) (*Opacity Project*) donnant les opacités des principaux éléments chimiques pour des températures supérieures à 6000 Kelvin. Ou encore Alexander & Ferguson (1994) et plus récement, Ferguson et al. (2005) pour les températures inférieures.

Le transport d'énergie par phénomène convectif est bien moins compris. Il n'existe pas à proprement parler de théorie satisfaisante du transport par convection et ces mécanismes intrinsèques constituent un problème d'actualité dans tous les domaines de la physique des milieux continus. Ajoutons que la modélisation d'une étoile se fait à une dimension dans la très grande majorité des cas car les calculs sont longs ; or il est dans ce cas difficile de décrire la convection que l'on observe pourtant fort bien à la surface du Soleil. Bien que divers formalismes pour traiter la convection existent, on en citera deux qui sont régulièrement usités lors de la modélisation de la structure interne d'une étoile : la théorie du mélange (MLT) et la théorie de Canuto-Goldman-Mazzitelli (CGM).

MLT. La théorie de la longueur de mélange est l'une des plus vielles description phénoménologique du transport convectif. Elle suppose que les mélanges ne se font qu'à une seule échelle convective l, correspondant à la taille caractéristique des cellules convectives. Elle est exprimée en fonction de la hauteur de pression $H_p = -\frac{P}{\rho g}$, P étant la pression totale du fluide (pressions thermique, radiative et turbulente) : $l = \alpha H_p$ (Böhm-Vitense 1958). Typiquement, dans un modèle solaire, $\alpha_{MLT} \approx 1.7$.

CGM. Ce formalisme est plus complet que la théorie de la longueur de mélange puisqu'il considère toute une gamme d'échelles turbulentes (Canuto et al. 1996). Typiquement, dans un modèle solaire $\alpha_{CGM} \approx 0.7$.

La seconde relation de fermeture est donnée par l'*équation d'état*. Elle établit une relation entre pression, densité et température du gaz pour une mixture chimique donnée. C'est une manière formelle de décrire la dépendance entre les variables d'états thermodynamiques indépendamment du temps. Par exemple, un gaz peut être détendu et comprimé plusieurs fois de suite, sans qu'il en garde mémoire. Il n'est pas nécessaire d'avoir l'historique des états passés pour en déduire

l'état présent (effet de mémoire faible). Une équation d'état commune est l'équation d'état d'un gaz parfait : $PV = nRT$.

Plusieurs équations d'état sont utilisées dans la communauté astérosismologique lorsqu'il s'agit d'effectuer des modélisations numériques stellaires. La plupart du temps, le calcul utilise des quantités numériques tabulées. Nous ne décrirons ici que deux équations d'états : MHD (initiales des 3 auteurs qui en sont à l'origine) et OPAL.

MHD. Cette description a été développée par Mihalas et al. (1988), Hummer & Mihalas (1988) et Daeppen et al. (1988) dans le cadre de l'*Opacity Project*. Cette approche prend en compte les atomes et les ions du plasma d'un point de vue chimique. Le phénomène d'ionisation y est donc vu comme une simple réaction chimique.

OPAL. L'équation d'état OPAL a été développée plus récemment par Rogers & Iglesias (1996) et s'est imposée comme un standard. En effet, les données sismiques solaires sont mieux reproduites avec OPAL qu'avec MHD. L'équation est définie *via* une approche de physique statistique (ensemble grand canonique) dans laquelle électrons et noyaux atomiques sont les constituants du plasma. Les populations des différentes espèces et leur état sont donc entièrement déterminés par le modèle statistique utilisé.

1.2.3 Pulsations stellaires et bestiaire d'ondes

Les équations précédentes décrivent la structure interne d'une étoile à l'équilibre. Les observations montrant clairement le caractère oscillant des étoiles. Pour exprimer ces oscillations, l'approche classique consiste à effectuer quelques hypothèses simplificatrices :

- les oscillations sont de faible amplitude au regard des dimensions de l'astre. Cette hypothèse est pleinement justifiée dans les étoiles de type solaire car les oscillations sont de quelques centaines de mètres, soit de quelques fractions à peine du rayon de l'étoile (rapport $\delta R/R$). On peut alors utiliser un développement perturbatif sur les équations de l'état d'équilibre statique et les linéariser. Pour chaque variable X on effectue le changement de variable,

$$X = X_0 + \delta X. \tag{1.8}$$

- on suppose les oscillations adiabatiques. Cette hypothèse est valide dans les zones radiatives mais ne l'est plus dans les zones convectives. Grâce à cette hypothèse, l'équation de conservation de l'énergie est simplifiée

$$\frac{1}{p}\frac{dp}{dt} = \Gamma_1 \frac{1}{\rho}\frac{d\rho}{dt}, \tag{1.9}$$

  avec Γ_1 l'exposant adiabatique défini par : $\Gamma_1 = \frac{dln\,p}{d\rho}|_{entropie=cte}$.
- on supposera aussi l'étoile sphérique. Tout effet perturbatif qui briserait la symétrie (rotation, champ magnétique, ...) est donc négligé. La symétrie sphérique implique une indépendance à la colatitude θ et à la longitude φ : $X_0(\boldsymbol{r}, t) = X_0(r, t)$.
- l'équilibre est stationnaire : $X_0(\boldsymbol{r}, t) = X_0(r)$ et $\frac{dX_0}{dt} = 0$. Cela vaut aussi pour la composition chimique de l'étoile.

Ces hypothèses nous permettent d'obtenir un nouveau système d'équations pour les variables δp, $\delta \boldsymbol{r}$, $\delta\Phi$. L'indépendance entre variables spatiales et temporelle (équilibre stationnaire) nous permet d'affirmer que toute quantité de la forme $\delta X(\boldsymbol{r}, t) = \delta X(\boldsymbol{r})\mathrm{e}^{-i\omega t}$ est solution du système. Enfin on peut aussi montrer qu'il y a séparation des variables d'espace dans la base $(\boldsymbol{e_r}, \boldsymbol{e_\theta}, \boldsymbol{e_\varphi})$.

Ainsi δp, $\delta\Phi$ et la composante radiale de déplacement $\delta\xi_r$ tel que [5] $\boldsymbol{\delta r} = \delta\xi_r e_r + \delta\xi_\perp = \delta\xi_r e_r + \delta\xi_\theta e_\theta + \delta\xi_\varphi e_\varphi$ sont des fonctions propres de l'opérateur définis par les harmoniques sphériques Y_l^m,

$$Y_l^m(\theta,\varphi) = (-1)^m c_{lm} P_l^m \cos\theta \mathrm{e}^{im\varphi}, \tag{1.10}$$

où P_l^m est la famille des polynômes de Legendre et c_{lm} des constantes de normalisation choisies telles que l'intégrale de $|Y_l^m|^2$ sur la sphère unité soit égale à un. La valeur propre associée étant $l(l+1)$, on a,

$$\Delta_\perp Y_l^m = Y_l^m(\theta,\varphi)\frac{l(l+1)}{r^2}. \tag{1.11}$$

Au final les solutions du système d'équations perturbées δX (pour $\delta X = \{\delta\xi_r, \delta p, \delta\Phi\}$) s'écrivent

$$\delta X(r,\theta,\varphi) = X_{\alpha lm}(r) Y_l^m(\theta,\varphi)\mathrm{e}^{i\omega_{\alpha lm}t}, \tag{1.12}$$

où $\omega_{\alpha lm} = 2\pi\nu_{\alpha lm}$ est la pulsation des modes, chacun d'eux étant caractérisé par le triptyque (α, l, m). α est un scalaire caractérisant les solutions selon le rayon. l et m, les solutions latitudinales et longitudinales.

Le système est linéaire et la solution générale est une combinaison linéaire de l'Eq.1.12,

$$\delta X^{tot}(r,\theta,\varphi) = \sum_{l=0}^{+\infty}\sum_{m=-l}^{m=+l} X_{\alpha lm}(r) Y_l^m(\theta,\varphi)\mathrm{e}^{i\omega_{\alpha lm}t} \tag{1.13}$$

Enfin, la composante perpendiculaire de l'équation de quantité de mouvement s'écrit [6]

$$\rho_0\frac{\partial^2\xi_\perp}{\partial t^2} = -\nabla_\perp\delta p + \rho_0\nabla_\perp\delta\Phi, \tag{1.14}$$

ce qui nous permet d'exprimer le déplacement perpendiculaire $\xi_\perp$.

Au final on peut formuler une expression du déplacement total,

$$\delta\boldsymbol{r} = \left\{ \begin{array}{c} \xi_r(r)\, Y_l^m(\theta,\varphi)\, \boldsymbol{e_r} \\ \xi_\perp(r)\, \frac{\partial Y_l^m(\theta,\varphi)}{\partial\theta}\, \mathrm{e}^{i\omega_{\alpha lm}t}\, \boldsymbol{e_\theta} \\ \xi_\perp(r)\frac{1}{\sin\theta}\frac{\partial Y_l^m(\theta,\varphi)}{\partial\varphi}\, \mathrm{e}^{i\omega_{\alpha lm}t}\, \boldsymbol{e_\varphi} \end{array} \right\} \tag{1.15}$$

La séparation des variables d'espace et de temps permet de résoudre le système de 4 équations (Eq.1.2 à 1.5) et d'en déduire 3 équations différentielles en r (les dépendances temporelles, longitudinale et latitudinale se simplifiant), permettant de déterminer le déplacement radial $\xi_r(r)$, la pression $p(r)$ et le potentiel gravitationnel $\Phi(r)$ de l'étoile. La perturbation du déplacement perpendiculaire $\xi_\perp(r)$ et de la densité peut alors être déduite des Eq.1.2 et 1.9. Le problème étant linéaire, les solutions sont donc connues à une constante près et pour satisfaire les conditions aux limites (au coeur et en surface) il nous faut ajuster ω. On se ramène alors à un problème aux valeurs propres dont les solutions ne seront non nulles que pour un ensemble discret de pulsation $\omega_{n,l,m}$, caractérisée par un ordre radial n discret en remplacement de α.

5. Nous sommes en coordonnées sphériques. Dans ce cas la composante perpendiculaire à e_r du Laplacien s'écrit, $\Delta_\perp f = \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta\frac{\partial f}{\partial\theta}) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 f}{\partial\varphi^2}$.

6. $\nabla_\perp f = \frac{1}{r\sin\theta}\frac{\partial f}{\partial\theta}e_\theta + \frac{1}{r\sin\theta}\frac{\partial f}{\partial\varphi}e_\varphi$

1.2.4 Propriétés des oscillations : approximation de Cowling

Une résolution numérique du système d'équations présenté précédemment (Eq.1.2 à 1.5) peut être effectuée pour obtenir une modélisation de la structure interne. Une détermination des fréquences d'oscillations relativement bonne peut alors en être déduite en supposant entre autre la sphéricité de l'étoile et l'adiabacité des oscillations.

Cependant, pour comprendre de manière plus globale les propriétés des modes de pulsations stellaires, des résolutions analytiques moyennant des hypothèse fortes suffisent. Pour cela nous allons effectuer l'approximation de Cowling (Cowling 1941) qui consiste à négliger la perturbation du potentiel gravitationnel. Cette approximation est valide pour les modes de haut ordre n ou pour les modes de haut degré l. Les mesures que nous avons actuellement à partir des satellites CoRoT et Kepler sont du type photométrique non résolues. Les modes observés ne sont donc pas de haut degré. Cependant, dans les étoiles de type solaire de la séquence principale, l'ordre radial est suffisamment élevé pour que l'approximation soit valide ($n >> l$).

En négligeant $\delta\Phi$, on obtient deux équations différentielles qui peuvent être combinées en une seule en définissant $\psi(r) = c^2\rho_0^{1/2}\nabla.\boldsymbol{r}$ (Deubner & Gough 1984),

$$\frac{d^2\psi(r)}{dr^2} + K(r)\psi(r) = 0, \tag{1.16}$$

On reconnait l'équation générale d'une onde, avec,

$$K(r) = \frac{1}{c^2}\left[S_l^2(\frac{N^2}{\omega^2} - 1) + \omega^2 - \omega_{ac}^2\right] \tag{1.17}$$

$$K(r) = \frac{\omega^2}{c^2}\left(1 - \frac{\omega_+^2}{\omega^2}\right)\left(1 - \frac{\omega_-^2}{\omega^2}\right) \tag{1.18}$$

Le milieu étant supposé adiabatique on a l'approximation,

$$\psi(r) = \frac{c^2\rho_0^{1/2}}{\Gamma_{1,0}}\left(\frac{p_1}{p_0} - \frac{\rho_0 g_0}{p_0}\xi_r\right) \tag{1.19}$$

Détaillons les termes de l'équation :

- g_0 est la gravité à l'équilibre hydrostatique,

$$g_0 = -\rho_0\frac{dp_0}{dr}. \tag{1.20}$$

- $\Gamma_{1,0}$ représente l'exposant adiabatique à l'équilibre
- c est la vitesse du son adiabatique définie par

$$c^2 = \frac{\Gamma_{1,0}p_0}{\rho_0}. \tag{1.21}$$

- N est la fréquence de Brunt-Vaisala,

$$N^2 = g_0\left(\frac{1}{\Gamma_{1,0}}\frac{d\,ln\,p_0}{dr} - \frac{d\,ln\,\rho_0}{dr}\right) = -g_0\left(\frac{g_0}{c^2} - \frac{1}{\rho_0}\frac{d\rho_0}{dr}\right). \tag{1.22}$$

C'est la fréquence à laquelle une "bulle" de gaz oscille verticalement autour d'une position d'equilibre, dans une atmosphère stable vis à vis des mouvements verticaux.

- S_l correspond à la fréquence de Lamb,

$$S_l^2 = \frac{l(l+1)c^2}{r^2} \tag{1.23}$$

- ω_{ac} la fréquence de coupure accoustique,

$$\omega_{ac} = \frac{c^2}{4H^2}\left(1 - 2\frac{dH}{dr}\right); \tag{1.24}$$

$H = -\frac{d\,ln\rho_0}{dr}$ désigne l'echelle caractéristique de hauteur pour la densité.
- ω_-, ω_+ sont définies à partir des fréquences précédentes par :

$$\omega_\pm^2 = \frac{1}{2}\left(S_l^2 + \omega_{ac}^2\right) \pm \left[\frac{1}{4}\left(S_l^2 + \omega_{ac}^2\right)^2 - N^2 S_l^2\right] \tag{1.25}$$

Le comportement des modes est déterminé par les variations avec r des fréquences caractéristiques N, S_l, ω_{ac} et $\omega_\pm$. Les Fig.1.6 et Fig.1.7 sont des diagrammes de propagation permettant une visualisation du comportement de ces fréquences caractéristiques en fonction de quelques modes pour un modèle solaire standard. On y distingue les zones de piégeage d'onde dans l'étoile. La base de la zone convective solaire est située à environ $r/R_\odot = 0.7$. N prend une valeur imaginaire quasi-nulle dans les zones convectives, mais devient positif dans les zones radiatives. En conséquence après un processus d'inversion ou d'inférence, c'est un bon indicateur de la position des transitions entre zones radiatives et convectives au sein d'une étoile. S_l une fonction décroissante du coeur jusqu'à la surface. ω_{ac} devient significative uniquement proche de la surface de l'étoile (grandeur non représentée).

Le comportement local de $\psi(r)$ dépend du signe de $K(r)$.

a. Si $K(r)$ est positif, $\psi(r)$ est une fonction localement oscillante. Cette condition est remplie si,
   i. $\omega^2 > \omega_-^2$ **et** $\omega^2 > \omega_+^2$ **ou**,
   ii. $\omega^2 < \omega_-^2$ **et** $\omega^2 < \omega_+^2$
b. Si $K(r)$ est négatif alors les solutions sont des exponentielles croissantes ou décroissantes. Ceci se produit tant que les conditions du point (a) ne sont pas vérifiées.

En conclusion, pour un mode propre donné, il existe un domaine dans lequel la solution est oscillante. Les points (i) et (ii) correspondent respectivement à des solutions pour lesquelles les modes se comportent comme des *modes p* (à haute fréquence) et à des solutions pour lesquelles les modes se comportent comme des *modes g* (à basse fréquence). Dans les deux cas, force de pression et de gravité agissent. Mais à suffisamment haut ordre radial, le comportement asymptotique nous amènera à pouvoir négliger l'une ou l'autre des forces.

1.2.5 Modes p, modes g et modes mixtes

Les modes p et modes g sont les deux principales familles de modes. On parle de mode p lorsque la fréquence d'oscillation est très supérieure à la fréquence de Brunt-Vaisala : $\omega_{n,l,m} >> N$. Dans ce cas, $K(r)$ est approximable par,

$$K(r) \simeq (\omega^2 - S_l^2 - \omega_{ac}^2)/c^2. \tag{1.26}$$

La dynamique des modes est alors dominée par la variation de la vitesse du son en fonction de r : on parle de *modes acoustiques*. La force de rappel étant dominée par la pression leur vaut aussi l'appellation de *modes p*.

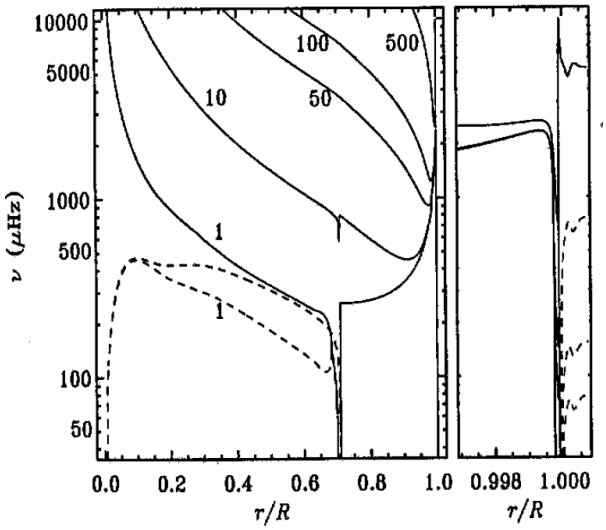


FIGURE 1.6 – Fréquence caractéristiques $\nu_+ = \omega_+/2\pi$ (ligne pleine) $\nu_- = \omega_-/2\pi$ (ligne en pointillés) pour un modèle solaire dans un diagramme de propagation. Les chiffres en vis à des vis des courbes indiquent l'ordre l du mode. La figure de droite est un grossissement au voisinage de la surface du Soleil. Diagramme tiré de Christensen-Dalsgaard (1997).

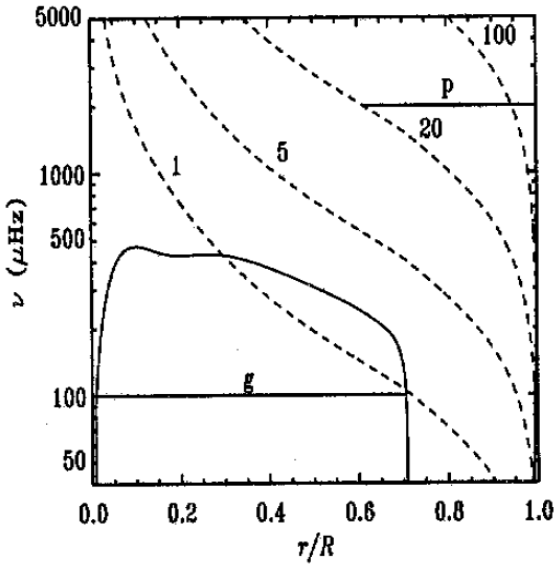


FIGURE 1.7 – Fréquence de Brunt-Vaisala (ligne pleine) et de Lamb (pointillés indéxé par l). Les *modes g* sont confinés dans la zone radiative solaire (le trait plein indexé "g" indique la cavité d'un mode de fréquence 100 μHz) tandis que les *modes p* se propagent à différentes profondeurs (le trait plein indexé "p" indique la cavité d'un mode $l = 20$ à une fréquence de 2000 μHz. Certains modes de plus bas degré peuvent avoir des propriétés mixtes. Dans le diagramme présenté, c'est par exemple le cas des modes $l = 1$ qui, pour une fréquence comprise entre 100 μHz et 400 μHz et pour $0.3 \lesssim R \lesssim 0.7$ se comporte comme un mode g, et comme un mode p ailleurs. En comparant cette figure avec la précédente, on remarquera que $\omega_+ \simeq S_l$ et que $\omega_- \simeq N$. Diagramme tiré de Christensen-Dalsgaard (1997).

La présence d'un gradient de température tout le long du rayon de l'étoile induit une réfraction continue des modes. On peut montrer que le point d'inflexion de l'onde est fonction de son degré, de sa pulsation et de la vitesse du son local,

$$r_t^2 = \frac{l(l+1)c^2(r_t)}{\omega^2}. \tag{1.27}$$

Tassoul (1980) a montré qu'en première approximation, la fréquence de pulsation de ces modes est une fonction linéaire de n et l,

$$\nu(n,l) = (n+\alpha+\frac{l}{2})\Delta\nu - A\frac{l(l+1)+\delta}{n+l/2+\alpha} \tag{1.28}$$

$$= (n+\epsilon+\frac{l}{2})\Delta\nu - l(l+1)D_0. \tag{1.29}$$

Dans cette relation, α ou ϵ sont des termes sensibles aux propriétés de surface de l'étoile. $\Delta\nu$ est appellé *grande séparation*, définie comme l'inverse du temps de trajet d'une onde accoustique, partant de la surface, passant par le coeur, puis retournant en surface,

$$\Delta\nu^{-1} = 2\int_0^R \frac{dr}{c(r)}. \tag{1.30}$$

Cette grandeur est très sensible aux régions où la vitesse du son est faible, c'est à dire essentiellement en surface. A est une grandeur sensible à l'intérieur stellaire. Elle fait intervenir la dérivée de la vitesse du son et sa valeur en surface,

$$A = \frac{1}{4\pi^2}\left[\frac{c(R_\odot)}{R_\odot} - \int_0^{R_\odot}\frac{dc}{dr}\frac{dr}{r}\right]. \tag{1.31}$$

Enfin, δ est une grandeur sensible aux propriétés internes à l'équilibre. Il est plus commun dans la pratique d'utiliser la formulation de l'Eq.1.29 plutôt que celle de l'Eq.1.28. La loi asymptotique permet de mettre en évidence la distribution régulière des modes p dans le spectre de puissance. Elle montre aussi que les modes de même parité ($l = 0$ et $l = 2$ par exemple) se regroupent par paires approximativement équidistantes dont la distance moyenne est obtenue par différence. On a par exemple,

$$\delta\nu_{01}^{(0)}(n) = \frac{\nu(n,0)+\nu(n+1,0)}{2} - \nu(n,1) = 2D_0 \tag{1.32}$$

$$\delta\nu_{01}^{(1)}(n) = -\frac{\nu(n,1)+\nu(n+1,1)}{2} + \nu(n+1,0) = 2D_0 \tag{1.33}$$

$$\delta\nu_{02}(n) = \nu(n+1,0) - \nu(n+1,2) = 6D_0 \tag{1.34}$$

$$\delta\nu_{03}(n) = \nu(n+1,1) - \nu(n,3) = 10D_0 \tag{1.35}$$

La relation de proportionnalité entre n et ν implique que le point tournant est une fonction du rapport l/n. Les modes de bas degré se propagent en profondeur dans l'étoile tandis que les modes de plus haut degré restent confinés en surface.

Pour les modes g, les points tournants sont déterminés par la condition $\omega = N$. Comme montré Fig.1.12, dans le cas solaire, cette condition impose qu'il y ait deux points tournants, l'un au voisinage du coeur, l'autre un peu en dessous de la base de la zone convective. Pour ces modes, à relativement haute fréquence, on a $\omega^2 << S_l^2$. Dans ces conditions, $K(r)$ peut être approximé par,

$$K(r) \simeq \frac{(N^2-\omega^2)l(l+1)}{r^2\omega^2}. \tag{1.36}$$

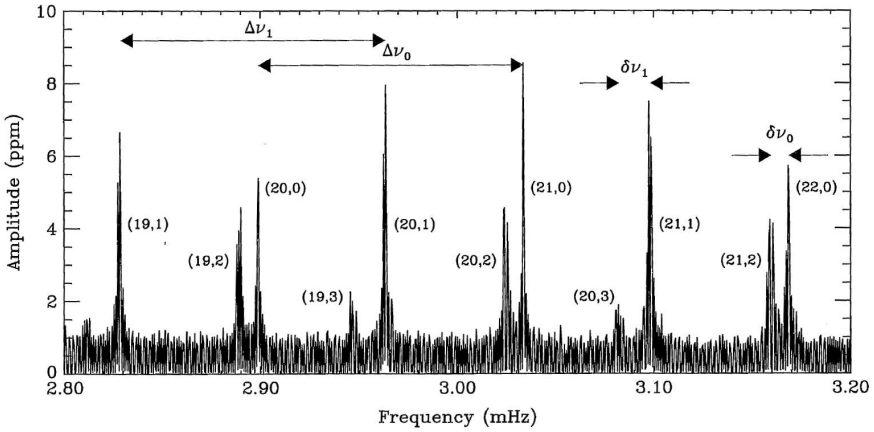


FIGURE 1.8 – Image montrant la répartition quasi-régulière des modes dans le spectre de puissance. Ici ce sont des observations en vitesse faite par l'instrument GOLF à bord de SoHo. On y voit les grandes séparation pour le $l = 0$ ($\Delta\nu_0$) et pour le $l = 1$ ($\Delta\nu_1$). Les petites séparations entre le $l = 2$ et le $l = 0$ ($\delta\nu_0$) et aussi entre le $l = 3$ et le $l = 1$ ($\delta\nu_1$) sont aussi visibles. Figure tirée de Aerts et al. (2010).

Cette fois ci, la dynamique des modes est dominée par les variations de N selon r. Ils sont stables dans les régions radiatives d'une étoile, évanescents dans les zones convectives. Tout comme pour les modes p, il existe une relation asymptotique décrivant leurs répartition approximative en période (*i.e.* dans le temps et non plus en fréquence),

$$\Pi(n, l) = \frac{\Pi_0}{\sqrt{(l(l+1))}}(n + \epsilon) \tag{1.37}$$

Π_0 est une fonction de la fréquence de Brunt-Vaisala,

$$\Pi_0^{-1} = \frac{1}{2\pi^2} \int_{cavité} \frac{N}{r} dr. \tag{1.38}$$

L'intégrale est calculée dans la cavité dans laquelle les modes g se propagent. Les modes g sont très sensibles aux propriétés au coeur d'une étoile de type solaire. Même s'ils sont profondément amortis une fois arrivés en surface, on a longtemps pensé que les durées d'observation extrêmement longues de SoHo combinées avec son extrême sensibilité permettraient de les débusquer. Malgré leur recherche active souvent ponctuée de fausse alertes, ils n'ont pu être mis en évidence de manière non ambigüe [7]. Ce n'est pas le cas pour d'autres types d'étoiles. Dans les naines blanches par exemple, leur analyse a permis de mesurer la stratification de leur atmosphère, mettant en évidence un coeur de carbone-oxygène et des couches superficielles d'hélium et d'hydrogène (Aerts et al. 2010).

D'une manière générale, la présence de discontinuités (le passage d'une zone convective à une zone radiative en est une) provoque des oscillations des fréquences autour des valeurs prédites par les lois asymptotiques des modes p et g. Ces oscillations de fréquence portent la signature de la nature de la discontinuité qu'il est intéressant d'extraire pour comprendre la structure interne de l'étoile (Monteiro et al. 1994).

7. Il est possible qu'ils aient été détecté tout récemment par Garcia et al. (2010)

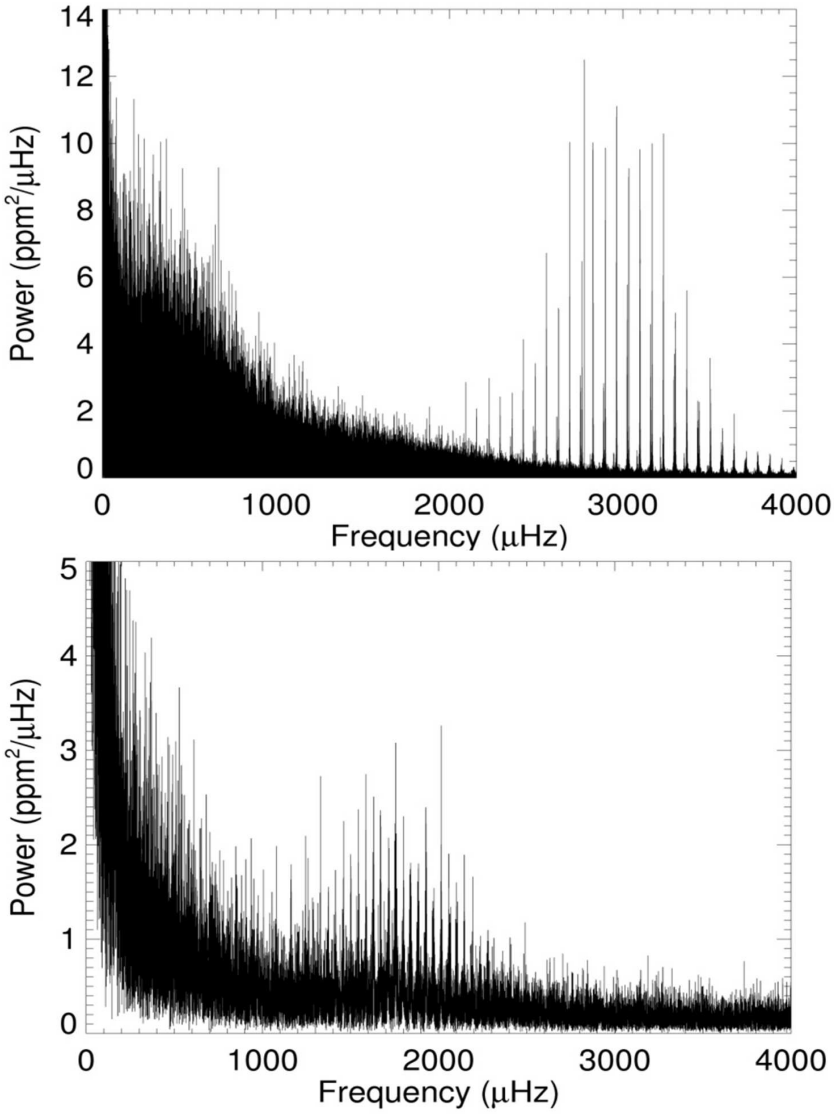


FIGURE 1.9 – Spectre de puissance du Soleil (haut) vue par VIRGO (SoHo) et d'une étoile de CoRoT HD49933 (bas). Dans les deux cas, la signature des modes et du bruit est clairement visible. On distingue aussi le motif régulier des modes, alternativement $l = 0 - 2$ et $l = 1 - 3$.

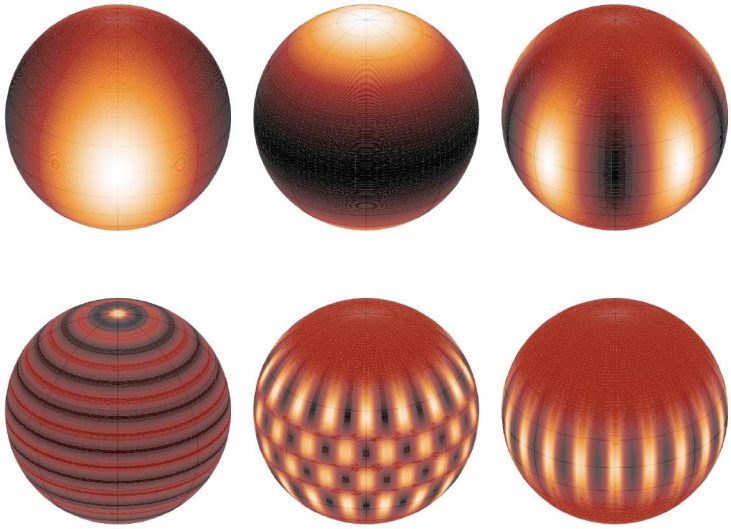


FIGURE 1.10 – Quelques exemples de modes de pulsation en fonction de m et l pour des modes de bas degré en haut, et de relativement plus haut degré en bas. Tirée de Tarrant (2010).

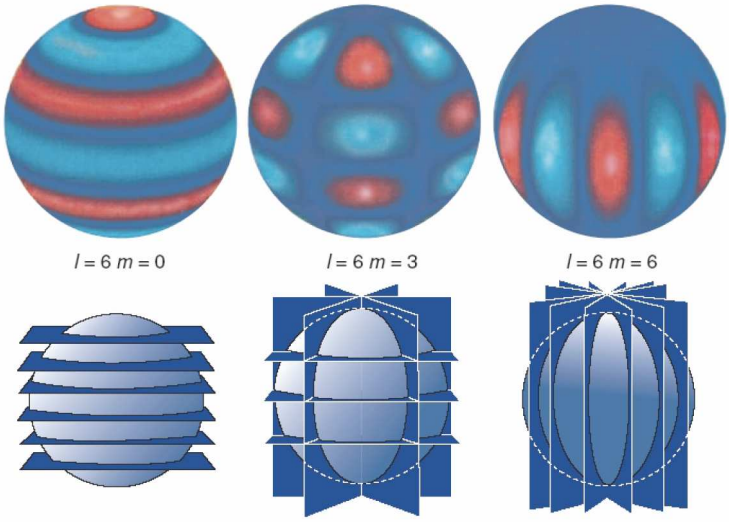


FIGURE 1.11 – Quelques exemples de modes de pulsation en fonction de m pour $l = 6$. Cet représentation montre que l représente le nombre de lignes nodales et m le nombre de lignes nodales coupant l'équateur.

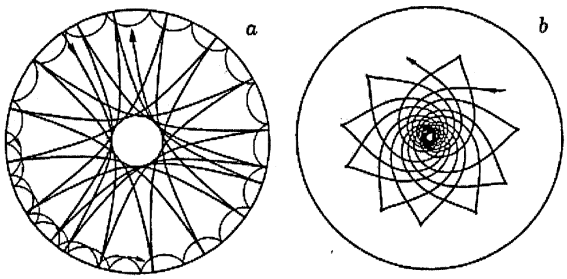


FIGURE 1.12 – Représentation schématique de la propagation d'ondes dans le Soleil pour les modes p se propageant dans tout l'intérieur stellaire (a) et les modes g, confinés dans la zone radiative (b).

Tiré de Christensen-Dalsgaard (1997).

Les modes mixtes ont la particularité de se comporter comme des modes p dans certaines régions de l'étoile et comme des modes g dans d'autres parties. Ils se manifestent essentiellement dans les étoiles évoluées ou en fin de séquence principale. Ils affectent essentiellement le $l = 1$ et dans une moindre mesure les degrés supérieurs car l'intensité du couplage entre la cavité p et la cavité g dépend de la distance les séparant. Or plus le degré d'un mode p est élevé, plus celui-ci est confiné en surface tandis que la zone de propagation d'un mode g est dictée par l'étendue de la zone radiative dans laquelle il est confiné. Si CoRoT a débusqué quelques étoiles porteuses de quelques modes mixtes, plus de la moitié des étoiles Kepler analysées en montrent. Les modes mixtes augmentent en nombre au fur et à mesure que l'étoile évolue. La loi asymptotique devient de moins en moins valide. La structure interne de l'étoile évolue vite dans le temps. En conséquence les propriétés des modes mixtes changent rapidement et permettent d'obtenir des contraintes très fortes sur l'age de l'étoile.

1.2.6 Influence de la rotation et visibilité des modes

Jusque là, nous avons négligé l'influence de toute perturbation brisant la symétrie sphérique. Pour les étoiles de type solaire, le champ magnétique n'a pas d'impact significatif contrairement à la rotation (Unno et al. 1989). Notons toutefois que des étoiles roAp semblent avoir un champ magnétique suffisamment intense pour que celui ci ait une influence plus importante que la rotation. Bien des astres ont des vitesses de rotations rapides, voire très rapides comparé au soleil.

Lorsque la symétrie sphérique est préservée, les fréquences propres des modes sont indépendantes de m : n et l suffisent à intégralement décrire la distribution modale. Une brisure de symétrie a pour effet de lever la dégénérescence des fréquences propres selon m. Pour chaque ordre radial de degré l, il existe alors $(2l + 1)$ modes de degrés m tel que $-l < m < +l$.

Physiquement, tenir compte de la rotation revient à tenir compte de la force de Coriolis et de la force centrifuge. Pour un rotateur lent en rotation solide, on peut montrer qu'au premier ordre (approximation de Ledoux) la nouvelle fréquence est fonction de la fréquence sans rotation et d'un terme perturbatif (appelé *splitting rotationnel*) fonction de la fréquence de rotation Ω,

$$\nu(n, l, m) = \nu(n, l) + m\Omega(1 - C_{nl}). \tag{1.39}$$

Dans le cas du soleil (rotation lente d'environ 28 jours), C_{nl} est très faible ($\leq 10^{-2}$) si bien que la distance entre deux ordres azimutaux m et $m \pm 1$ est très proche de la fréquence de rotation

stellaire : la rotation est dominée par la force de Coriolis. Notons toutefois que le Soleil n'est qu'une étoile parmi d'autres : il est possible d'avoir des étoiles de type solaire dont la période de rotation se chiffre en quelques jours voire quelques heures. En conséquence, les multiplets $m \neq 0$ des modes (n, l) peuvent par exemple être asymétriques par rapport au $m = 0$.

Pour des rotateurs rapides, la force centrifuge peut fortement déformer la structure sphérique (aplatissement des pôles important) si bien qu'il est nécessaire de faire intervenir le terme perturbatif du second ordre, proportionnel à $\frac{\Omega^2 R^3}{GM}$. Celui-ci correspond au rapport de la force centrifuge et de la force gravitationnelle à la surface (Saio 1981). Notons enfin que la rotation est dans la réalité une fonction de la profondeur r et de la colatitude θ, donc de n et de l.

En astérosismologie, un autre facteur important est l'angle formé entre l'axe de rotation d'une étoile et la ligne de visée de l'observateur. L'intensité d'un mode à la surface d'une étoile est une fonction des harmoniques sphériques et du temps,

$$I(\theta, \varphi, t) \propto Re\left\{I_0 Y_l^m(\theta, \varphi) \mathrm{e}^{i\omega t}\right\}. \tag{1.40}$$

Les étoiles étant en quasi-totalité des corps non résolus spatialement, l'intensité mesurée est l'intégrale sur toute sa surface visible (mesure en *disque intégré*),

$$I_m(t) \propto \int_{\Xi=\{disque\}} I(\theta, \varphi, t) d\Xi. \tag{1.41}$$

L'intensité dépend alors des modes l et m et de la section de disque visible par l'observateur. On peut distinguer deux types d'effet. Tout d'abord, par un effet de projection, l'intensité mesurée est plus ou moins sensible au centre de l'étoile plutôt qu'au bord. Par exemple, la profondeur d'atmosphère stellaire traversée par le rayonnement n'est pas la même au centre qu'au niveau du bord du disque de l'étoile. Ces effets de projection sont appelés *effet centre bord* et affectent l'intensité totale mesurée de chacun des degrés l (indépendamment de m). Il est utile de définir la hauteur totale des modes (n, l) en fonction de celle du $l = 0$ (celle des composantes m étant définie en fonction du l associé). Ainsi pour le Soleil dans le spectre de puissance, le $l = 1$ est 1.5 fois plus intense que le $l = 0$, le $l = 2$ est 0.53 plus faible, la hauteur du $l = 3$ est presque nulle (0.027) et celle des $l > 4$ négligeable. On ajoutera tout de même que la nature de l'atmosphère de l'étoile est soupçonnée de pouvoir sensiblement changer ces rapports, notamment pour les plus évoluées d'entre elles.

Deuxièmement, comme l'a montré Gizon & Solanki (2003), les multiplets m à l donné ne sont pas tous visibles en même temps. Les observations à partir du plan de l'écliptique du Soleil (angle d'inclinaison proche de 90°) ne permettent pas par exemple de voir les composantes $m = \pm 1$ du $l = 1$ en luminosité intégrée. Une représentation des effets angulaires pour les modes de bas degré qui nous intéressent lors de mesures astérosismiques est donnée Fig.1.13.

1.2.7 Sur les mécanismes d'excitation des modes

Comprendre le mécanisme d'excitation des modes est nécessaire pour déduire la forme du profil des modes. Deux principaux modèles ont été développés pour décrire la manière dont les modes sont excités.

Le premier, nommé *mécanisme* κ a été proposé initialement par Ando & Osaki (1975) et fait intervenir l'opacité du milieu.Toutefois ce mécanisme ne permet pas de reproduire correctement les amplitudes d'oscillations constatées sur le Soleil (amplitudes trop importante). Bien que ce mécanisme soit prépondérant dans d'autres étoiles ($\delta\, Scuti$ ou autres pulsateurs classiques par exemple), il est plus que probable qu'il ne le soit pas dans les étoiles de type solaires à enveloppe externe convective.

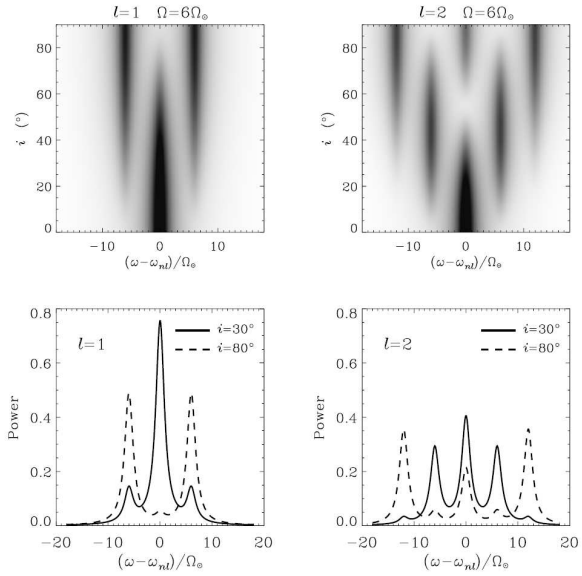


FIGURE 1.13 – Intensité relative des modes de bas degré dans le spectre de puissance en fonction de l'inclinaison. $i \simeq 90°$ correspond au cas d'observations solaires. Le degré de splitting rotationnel Ω est exprimé en unité de splitting solaire $\Omega_\odot$. Tiré de Gizon & Solanki (2003)

Le second mécanisme proposé par Goldreich & Keeley (1977) suppose que les modes stellaires sont excités stochastiquement par les mouvements de convection et amortis par des effets dynamiques et thermiques. En effet, pour reprendre notre référence favorite, les granules de surface solaires ont des temps de vie similaires aux périodes des modes et leur énergie cinétique est comparable à l'énergie des modes. Ajoutons que la prédiction du profil spectral des modes est compatible avec cette hypothèse. Ce profil est la conséquence des multiples réexcitations des ondes en surface, de phase et d'amplitude aléatoire. Enfin du fait du large spectre d'énergie offert par les phénomènes convectifs, ce mécanisme permet l'excitation d'un grand nombre de modes propres.

Dans ce modèle, chaque mode propre peut être représenté par un oscillateur amorti et aléatoirement réexcité par interaction avec la convection turbulente. L'amplitude de ces modes est donc solution de l'équation différentielle (Anderson et al. 1990),

$$\frac{1}{\omega_0^2}\frac{d^2}{dt^2}y(t) + \frac{2\pi\Gamma}{\omega_0^2}\frac{d}{dt}y(t) + y(t) = f(t), \tag{1.42}$$

où y représente le déplacement de surface, ω_0 la pulsation propre de l'oscillateur non amorti, Γ une constante relative au temps caractéristique d'amortissement et $f(t)$ une fonction stochastique décrivant la source d'excitation.

Une famille de solutions de l'équation précédente peut être formulée sous forme d'un produit de convolution,

$$y(t) = (h * f)(t) = \int_{-\infty}^{+\infty} h(u)f(t-u)du. \tag{1.43}$$

$h(u)$ est la réponse impulsionnelle du système (réponse à un Dirac). C'est elle que nous cherchons. La transformée de Fourier de l'Eq.1.43 s'exprime comme,

$$Y(\nu) = H(\nu)F(\nu), \tag{1.44}$$

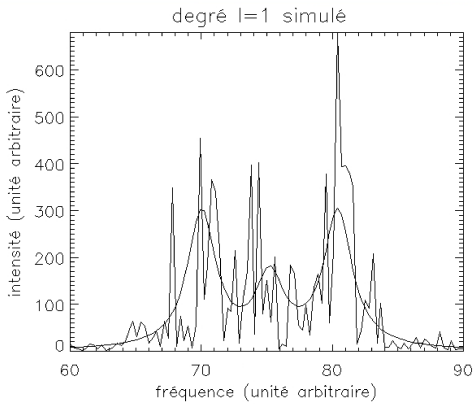


FIGURE 1.14 – Le profil théorique d'un mode de pulsation stellaire (ici un $l = 1$) superposé à son spectre de puissance (données synthétiques). On a supposé que l'axe de visé de l'observateur forme un angle de 60° avec l'axe de rotation de l'étoile.

dont les termes désignent respectivement les transformées de Fourier de $y(t)$, $h(t)$ et $f(t)$. En posant $f(t) = \delta(t - t_0)$ avec t_0 l'instant de l'impulsion, on calcul à partir de l'équation 1.42 la réponse à l'impulsion,

$$H(\nu) = \frac{\nu_0^2 \mathrm{e}^{2i\pi\nu t_0}}{\nu_0^2 - \nu^2 + i\Gamma\nu}. \tag{1.45}$$

Si on suppose que l'étoile n'est pas trop évoluée, alors la fréquence des pulsations est relativement élevé, (*i.e.* $\nu_0 \gtrsim 1000$ μHz). En supposant les largeurs solaires représentatives de celles des étoiles de type solaire, on peut alors supposer $\Gamma << \nu_0$. Le profil d'un mode dans le spectre de puissance s'en déduit aisément,

$$P(\nu) \propto |H(\nu)|^2 \tag{1.46}$$

$$\simeq \frac{A}{1 + \frac{(\nu-\nu_0)^2}{\Gamma^2/4}}. \tag{1.47}$$

Chaque mode de pulsation stellaire amortie, dont la source d'excitation est la convection turbulente est donc en théorie bien représenté par une lorentzienne dans le spectre de puissance.

On insistera sur le fait que l'excitation stochastique globale est le résultat d'un grand nombre d'excitations stochastiques de structures plus ou moins ponctuelles (l'ensemble des granules et structures de surfaces). En conséquence en vertu du théorème central-limite, le bruit aléatoire résultant est de nature gaussienne, ce profil est donc le profil moyen asymptotique (temps d'observation tendant vers l'infini). Comme nous l'expliquons Chap.2.1.2 le bruit gaussien dans l'espace temporel implique que la statistique du spectre de puissance suit une loi exponentielle décroissante (χ_2^2). La Fig.1.14 représente un exemple de mode excité stochastiquement.

Au final, le problème est linéaire et le profil moyen d'excitation des modes dans son intégralité est défini par,

$$S(\nu, \boldsymbol{\theta_S}) = \sum_{n=n_0}^{N_{\max}} \sum_{l=0}^{L_{\max}} \sum_{m=-l}^{+l} \frac{H(n,l,m)}{1 + 4(\frac{\nu-\nu(n,l)+m\nu_s}{\Gamma(n,l,m)})^2}. \tag{1.48}$$

Dans cette relation on a explicitement défini la dépendance du modèle aux différents paramètres des modes au travers du vecteur $\boldsymbol{\theta_S}$.

L'ajustement d'un tel modèle d'excitation dans le spectre de puissance nous permet d'obtenir des informations intéressantes sur l'amplitude, le temps de vie caractéristique et la fréquence des modes. Notons toutefois que les hypothèses formulées peuvent ne pas être tout à fait correctes dans certaines circonstances et peuvent amener à de légers biais statistiques. On fera par exemple remarquer que le profil des modes solaires n'est pas tout à fait lorentzien : il existe une légère asymétrie des modes dont l'origine est liée aux propriétés intrinsèques des modes propres d'oscillation et de la localisation de la source d'excitation. En tant que science récente, l'astérosismologie d'aujourd'hui appliquée aux étoiles de type solaire peu évoluées ne permet pas, à notre avis, des mesures suffisamment précises pour qu'il soit nécessaire de prendre en compte les éventuels effets d'asymétrie des modes. Lorsque Kepler nous prodiguera des observations longues de plusieurs années sur quelques cibles, il faudra probablement alors affiner les modèles du profil des modes. Par exemple, en recherchant une asymétrie du splitting ou une asymétrie du profil des modes.

1.2.8 Relation empirique sur les étoiles de type solaire

A partir du moment où l'on s'intéresse à des étoiles dont le comportement supposé est voisin de celui de notre Soleil (en terme de mécanisme d'excitation notamment), il est utile de rechercher des lois d'échelles empiriques sur les quelques grandeurs caractéristiques stellaires.

C'est à quoi s'est essayé la communauté astérosismique. Même si les lois d'échelles qui sont présentées dans ce qui suit doivent être considérées avec précaution lors d'études d'astres pris individuellement, elles permettent de déduire de manière assez fiable le comportement moyen d'étoiles montrant des oscillations de type solaire (ce qui inclut les géantes rouges).

Soit $A \equiv \delta L/L$ la variation de luminosité relative due aux pulsations. Les variations de volume induites par la détente/compression du plasma stellaire étant négligeables, on peut relier ces variations directement aux variations de température (gaz parfait) et donc de pression,

$$\frac{\delta L}{L} \propto \frac{\delta T}{T} \propto \frac{\delta \rho}{\rho}. \tag{1.49}$$

Moyennant quelques hypothèses supplémentaires (Kjeldsen & Bedding 1995) et prenant le Soleil comme référence, on peut montrer que la luminosité relative d'une étoile varie approximativement selon une fonction de sa température effective de surface T_{eff} et de sa masse M,

$$A = \frac{L/L_\odot}{(M/M_\odot)(T_{eff}/T_\odot)} A_\odot. \tag{1.50}$$

Samadi et al. (2007) ont proposé de réviser légèrement cette loi empirique en ajoutant un exposant e,

$$A = \left(\frac{L/L_\odot}{M/M_\odot}\right)^e \frac{1}{(T_{eff}/T_\odot)} A_\odot. \tag{1.51}$$

Selon les observations les plus récentes, l'exposant vaut environ $e = 0.7$.

Par des considérations similaires et en se servant des résultats précédents, Deheuvels et al. (2010) définit l'amplitude attendue des modes en fonction de l'amplitude des modes solaires et de la fréquence centrale des modes ν_{max},

$$A_{max} = A_\odot (T_{eff}/T_\odot^{1.95})(\nu_{max}/\nu_{max,\odot})^{-0.7}. \tag{1.52}$$

La fréquence centrale d'une étoile peut aussi être approximée par une loi d'echelle,

$$\nu_{max} = \frac{M/M_\odot}{(R/R_\odot)^2 \sqrt{T/T_\odot}} \nu_{max,\odot}. \tag{1.53}$$

On rappelle que pour le soleil, $A_\odot = 4,1 \pm 0.3$ ppm (amplitude approximative du $l = 0$), $\nu_{max,\odot} = 3100$ μ Hz et $T_\odot = 5770$ K.

A partir des données les plus récentes et comme discuté par Stello et al. (2009), la relation,

$$<\Delta\nu>_n = (0.24 \pm 0.05)\nu_{max}^{0.78\pm0.045} \tag{1.54}$$

reliant la grande séparation $\Delta\nu$ et la fréquence centrale de l'enveloppe des modes ν_{max}, a été mise en évidence.

Enfin, il est possible d'établir une estimation de la hauteur des modes, dont voici une démonstration. Supposons que le profil de l'enveloppe des modes est gaussien de fréquence centrale ν_{max} et de la largeur σ. Cette approximation est valide dans le cas solaire et le semble aussi dans la grande majorité des pulsateurs du même type que le Soleil. La puissance totale des modes s'écrit donc,

$$P_{env} = \int P_0 \mathrm{e}^{-\frac{1}{2}(\nu-\nu_{max})^2/\sigma^2} d\nu = \sqrt{2\pi} P_0 \sigma \, (ppm^2). \tag{1.55}$$

En vertu du principe de conservation de l'énergie, la puissance totale peut aussi s'écrire en fonction de la somme des puissances individuelles des modes. Pour un profil lorentzien on a,

$$P_{mode} = \int \frac{A}{1 + \frac{(\nu-\nu_0)^2}{\Gamma/4}} d\nu = \frac{\pi H \Gamma}{2} \, (ppm^2). \tag{1.56}$$

Donc la puissance totale,

$$P_{t,modes} = P_{env} = \sum_{n,l,m} \pi \Gamma_{n,l,m} H_{n,l,m}/2. \tag{1.57}$$

Les hauteurs et largeurs varient en théorie selon (n, l, m). Cependant pour les hauteurs, nous n'avons pas besoin de tenir compte de m puisque nous connaissons les puissances totales de chaque multiplet en fonction de celle du $l = 0$ (cf. Chap.1.2.6),

$$V_{l=1} = \frac{H_{l=1}}{H_{l=0}} = 1.5 \tag{1.58}$$

$$V_{l=2} = \frac{H_{l=1}}{H_{l=0}} = 0.53 \tag{1.59}$$

$$V_{l=3} = \frac{H_{l=1}}{H_{l=0}} = 0.027. \tag{1.60}$$

Les hauteurs tel que $l > 3$ sont négligeables en disque intégré. Supposons alors que le profil de hauteurs est à l'image du profil de l'enveloppe, gaussien [8],

$$H(\nu, l) = h_0 \mathrm{e}^{-\frac{1}{2}(\nu(n,l)-\nu_{max})^2/\sigma^2}. \tag{1.61}$$

En première approximation, les modes occupent des positions $\nu(n, l)$ données par la loi asymptotique. En supposant $D_0 << \Delta\nu$ et $\epsilon \cong 0$, on peut simplifier la distribution des fréquences en l'exprimant en fonction de $\Delta\nu/2$. Ceci nous permet de nous affranchir de la dépendance en l,

$$\nu(n) = n\Delta\nu/2. \tag{1.62}$$

8. Si l'enveloppe des modes est réellement gaussienne (ce qui n'est pas toujours vraie), ceci revient aussi à supposer la largeur individuelle des modes constante.

Nous pouvons nous permettre une tel approximation car nous ne cherchons qu'a connaitre l'énergie totale contenue dans le profil. En conséquence, compte essentiellement l'écart moyen entre modes successifs. Enfin, on supposera que nous ne connaissons que la largeur moyenne des modes. Ceci revient à négliger les variations de largeurs : $\Gamma_{n,l,m} = \Gamma$.

On peut donc approximer la puissance totale en ne tenant compte que des modes les plus intenses (pour les besoins de la démonstration on négligera les degrés supérieurs à 2),

$$P_{t,modes} \simeq \pi h_0 \Gamma \sum_n \mathrm{e}^{-\frac{(\nu(n,0)-\nu_{max})^2}{2\sigma^2}} + 1.5\mathrm{e}^{-\frac{(\nu(n,1)-\nu_{max})^2}{2\sigma^2}} + 0.53\mathrm{e}^{-\frac{(\nu(n,2)-\nu_{max})^2}{2\sigma^2}} \quad (1.63)$$

$$\simeq 3\pi h_0 \Gamma \sum_n \mathrm{e}^{-\frac{(\nu(n)-\nu_{max})^2}{2\sigma^2}} \quad (1.64)$$

Lorsque $\Delta\nu < \sigma$, on peut montrer que la somme de droite est donnée par,

$$\sum_n \mathrm{e}^{-\frac{(\nu(n)-\nu_{max})^2}{2\sigma^2}}|_{\Delta\nu<\sigma} = 2\sqrt{2\pi}\frac{\sigma}{\Delta\nu}. \quad (1.65)$$

Nous ne connaissons pas de cas dans lequel l'hypothèse précédente ne soit pas valide si bien qu'au final nous pouvons relier la hauteur de l'enveloppe à la hauteur des modes,

$$h_0 = \frac{P_0 \Delta\nu}{\pi \xi \Gamma}, \quad (1.66)$$

où $\xi \cong 3$, dépend du nombre de degré supposé significativement intense dans le spectre.

Toutes ces informations sont utiles pour définir un point de départ lors d'ajustement de modèle des modes sur les spectres de puissances. On les utilisera notamment Chap.5, chapitre dans lequel nous présentons un algorithme d'ajustement des modes automatiques.

Tout au long de ce chapitre, nous avons vu que l'étude des ondes sismiques nous permet d'extraire une information sur la structure interne des étoiles. Nous allons maintenant montrer sur la base d'approches statistiques, de quelle manière cette information peut être extraite.

Chapitre 2

Sur les méthodes d'analyse et de traitement de signal classiquement usitées en astérosismologie

Sommaire

2.1 Vraisemblance et approche MLE 31
2.1.1 Généralités 31
2.1.2 Fonction de vraisemblance d'un spectre de puissance 33
2.1.3 Comparaison de modèles avec une approche MLE 34
2.2 Approche Bayésienne simple : Approche MAP 35
2.2.1 Probabilité *a priori* et probabilité *a posteriori* 35
2.2.2 Application de l'approche Bayésienne au cas stellaire 36
2.2.3 Comparaison de modèles avec une approche MAP 37
2.3 Le calcul d'erreur avec des algorithmes de type MLE ou MAP . . 38

2.1 Vraisemblance et approche MLE

2.1.1 Généralités

L'approche *Maximum Likelihood Estimator* (MLE) est une approche statistique popularisée par R. Fisher au début du 20ème siècle et depuis lors, très largement utilisée dans le domaine des sciences. Nous n'allons qu'esquisser ses bases. Une présentation approfondie de la méthode peut être trouvée par exemple dans Kendall & Stuart (1967).

Commençons par un exemple simple. Supposons que nous disposions d'un ensemble de mesures $\boldsymbol{x} = \{x_0, x_1, ..., x_N\}$. Si ces quantités sont associées à une variable unique, chaque mesure peut être vue comme constituée d'une valeur "vraie" s, perturbée par un bruit (additif) aléatoire b, $x = s + b$. $\boldsymbol{x}$ est donc une *variable aléatoire* dont la *fréquence* (du point de vue fréquentiste) est s. Nous pouvons alors attribuer une fonction de densité de probabilité $\pi(\boldsymbol{x})$ à cet ensemble de mesure $\boldsymbol{x}$ dont les x_i en sont des échantillons.

Dans l'approche MLE, on suppose l'indépendance mutuelle des mesures x_i et nous pouvons

écrire la probabilité conjointe sachant que les mesures sont paramétrisées par la variable s comme,

$$\pi(\boldsymbol{x}|s) = \prod_{i}^{N} \pi(x_i|s) \tag{2.1}$$

Nous voyons que cette quantité ne correspond pas à la probabilité que la variable s soit en accord avec les données, mais l'inverse (probabilité que les données soit en accord avec une valeur de s fixé). Le résultat de l'équation 2.1 est appelé la *vraisemblance* (ou *likelihood* en anglais).

On fera remarquer que dans cette optique, le monde est clairement séparé en des composantes aléatoires dont le comportement suit des lois statistiques, et des valeurs "vraies", qui idéalement, au bout d'un nombre infini de mesures, serait parfaitement déterminées. Ceci est la caractéristique même d'une approche fréquentiste.

Au maximum de $\pi(\boldsymbol{x}|s)$, l'estimateur vraisemblance fournit la valeur la plus probable de s, objet de la mesure. Autrement dit, le maximum définit la position dans l'espace des paramètres qui s'accorde au mieux aux mesures.

En général les fonctions de vraisemblance sont assez complexes et insolubles analytiquement. Ceci est encore plus vrai pour des statistiques non gaussiennes, éventuellement multimodales, auxquelles on est couramment confronté en astérosismologie. Les approches analytiques sont alors impossibles et il est nécessaire de trouver le maximum de la vraisemblance par des méthodes numériques.

Des problèmes numériques nous incitent à plutôt travailler avec le *log-vraisemblance* et pour des raisons algorithmiques [1], il est plus simple de minimiser une fonction que de la maximiser. C'est pourquoi on travaille toujours avec l'opposé du log-vraisemblance ($-$ *log-vraisemblance*). On appelle aussi cette quantité la fonction erreur car à son minimum elle définit le lieu dans l'espace des paramètres pour lequel le désaccord entre mesures et modèle est le plus faible.

En généralisant l'expression 2.1 au cas multidimensionnel, nous avons

$$l(\boldsymbol{x}|\boldsymbol{\theta}) = -ln(\pi(\boldsymbol{x}|\boldsymbol{\theta})) = -\sum_{i}^{N} ln(\pi(x_i|\boldsymbol{\theta})) \tag{2.2}$$

$\boldsymbol{\theta}$ étant un vecteur de paramètres (pour un modèle donné). Minimiser l'Eq.2.2 revient à résoudre un système à P équations aux dérivées partielles, P étant la dimension de $\boldsymbol{\theta}$:

$$\forall k \in [1, P],\ \frac{\partial l(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_k} = 0 \tag{2.3}$$

Cet estimateur est théoriquement asymptotiquement non biaisé et de minimum variance (l'estimateur ne fourni pas d'erreurs excessivement grandes), c'est à dire,

$$lim_{N\to\infty} E(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta} \tag{2.4}$$

$$lim_{N\to\infty} Var(\hat{\boldsymbol{\theta}}) = Var[\boldsymbol{\theta}] \tag{2.5}$$

$\hat{\boldsymbol{\theta}}$ étant l'estimé de $\boldsymbol{\theta}$. Vu la possible complexité de la fonction vraisemblance, il est courant de recourir à des méthodes numériques pour résoudre l'Eq. 2.3.

1. Les algorithmes développées au cours de l'histoire sont basés sur la recherche du minimum d'une fonction erreur, non pas du maximum de probabilité.

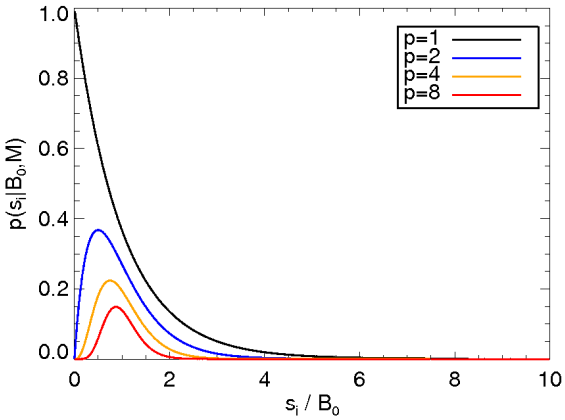


FIGURE 2.1 – Distribution du χ^2_{2p} en fonction du rapport signal à bruit S_i/B_0, pour un problème simple (cas ou le modèle est un bruit blanc). On remarquera que pour $p = 1$ la distribution est exponentielle. p s'accroissant, celle ci se rapproche de plus en plus d'une gaussienne.

2.1.2 Fonction de vraisemblance d'un spectre de puissance

L'approche MLE a été très largement utilisée en héliosismologie depuis que Duvall et Harvey l'on introduite en 1986 (Duvall & Harvey 1986).

Pour calculer le terme de vraisemblance, la première étape consiste à déterminer à quelle statistique répond chaque échantillon de notre mesure. Dans la grande majorité des cas en physique, la statistique est gaussienne car les mesures sont effectuées sur des systèmes physiques macroscopiques (théorème de la limite centrale).

C'est le cas en astéroseismologie des mesures en intensité. Chap.1.2.7 nous avions dit que chaque mode est essentiellement caractérisé par une fréquence propre de pulsation dont la fonction excitatrice est stochastique à cause de la turbulence atmospherique stellaire. On en a conclu que le bruit d'excitation des modes était gaussien car l'excitation se fait sur une large gamme spectrale, à différentes échelles spatiales et temporelle, dont les effets moyens engendrent un bruit gaussien. Si le bruit de la partie réelle et imaginaire d'un spectre de Fourrier $s(\nu_i)$ de chaque fréquence ν_i est gaussien alors la distribution statistique du spectre de puissance $S(\nu_i)$ correspondante est définie comme,

$$S(\nu_i) = Re[s(\nu_i)]^2 + Im[s(\nu_i)]^2, \tag{2.6}$$

est une distribution exponentielle (Scargle 1982) : nous sommons deux distributions gaussiennes. Cette distribution est plus communément appelée statistique du "khi deux à deux degrés de liberté" (χ^2_2).

Plus généralement on peut démontrer que la statistique de chaque point de la somme (ou moyenne) de p spectres de puissance est un χ^2 à $2p$ degrés de liberté (χ^2_{2p}). La statistique d'un point du spectre de puissance $S(\nu_i)$ s'écrit alors (Appourchaux 2003),

$$\pi(S(\nu_i)|\boldsymbol{\theta}, M) = p^{p-1}\frac{S(\nu_i)^{p-1}}{\Gamma(p)M(\nu_i|\boldsymbol{\theta})^p}e^{-p\frac{S(\nu_i)}{M(\nu_i|\boldsymbol{\theta})}}. \tag{2.7}$$

Où $\Gamma(p) = (p-1)!$ est la *fonction Gamma*.

$\pi(S(\nu_i)|\boldsymbol{\theta}, M)$ représente la probabilité que la valeur du spectre $S(\nu_i)$ s'éloigne du comportement moyen $M(\nu_i|\boldsymbol{\theta})$ de ce dernier, défini par le modèle M et paramétrisé par le vecteur $\boldsymbol{\theta}$.

Prenons un exemple simple : un spectre de puissance composé uniquement d'un bruit blanc B_0. Dans ce cas, le modèle est simplement, $M(\nu|B_0) = B_0$ et $\pi(S(\nu_i)|B_0, M) = \pi(S_i|B_0, M)$ représente la probabilité que $S(\nu_i)$ appartienne au bruit B_0 en ν_i. La relation 2.7 est représentée pour quelques valeurs de p, Fig.2.1, dans ce cas particulier.

Finalement, sur la base des équations 2.1 et 2.7, la vraisemblance pour un spectre de puissance constitué de N points, avec un nombre p quelconque de degrés de liberté s'écrira,

$$\pi(S(\boldsymbol{\nu})|\boldsymbol{\theta}, M) = \frac{p^{p-1}}{\Gamma(p)} \prod_i^N \frac{S(\nu_i)^{p-1}}{M(\nu_i|\boldsymbol{\theta})^p} e^{-p\frac{S(\nu_i)}{M(\nu_i|\boldsymbol{\theta})}}. \tag{2.8}$$

Cette relation fondamentale est la base de toute approche statistique appliquée à un spectre de puissance. Elle sera continuellement utilisée, y compris dans les chapitres traitant de l'approche par maximum à posteriori ou d'échantillonnage MCMC. Par la suite, pour alléger l'écriture, on notera $\boldsymbol{y}$ les données $S(\boldsymbol{\nu})$.

2.1.3 Comparaison de modèles avec une approche MLE

Nous avons montré que la vraisemblance est construite sur la comparaison d'un modèle (ou représentation des données) avec les mesures/données. Qu'en est il de la comparaison entre modèles ? En effet, théoriquement une infinité d'ajustements peut être effectuée, chacun basé sur une représentation différente des mesures.

Dans la pratique, des considérations d'ordre théorique nous permettent d'avoir un nombre fini de modèles compatibles avec les mesures. Pour distinguer les différentes représentations possibles des données, il est alors préférable d'avoir un critère objectif de comparaison de modèle.

Le contexte statistique nous permet de formuler un tel critère. Comme nous l'avons dit, le maximum de vraisemblance est la position dans l'espace des paramètres où les données s'accordent au mieux au modèle utilisé. Si nous disposons donc de deux modèles M_1 et M_2, leur maximum de vraisemblance peuvent être comparés pour établir lequel est le plus cohérent avec les mesures.

Formellement, ceci s'effectue via le calcul du *odds ratio* ou coefficient concurrentiel en français. Celui ci est défini comme le rapport,

$$O_{M_1,M_2} = \frac{\arg\max\{\pi(\boldsymbol{y}|\boldsymbol{\theta_1}, M_1)\}}{\arg\max\{\pi(\boldsymbol{y}|\boldsymbol{\theta_2}, M_2)\}}. \tag{2.9}$$

On peut alors définir une probabilité relative de M_1 par rapport à M_2,

$$P(M_1|\boldsymbol{y}) = (1 + O_{M_1,M_2}^{-1})^{-1} \tag{2.10}$$

Typiquement, un coefficient concurrentiel de 3 correspond à une probabilité relative de 75% [2]. En général, les seuils de rejection/d'acceptation de modèles sont à 90% ou plus. En deçà, le modèle dominant est considéré comme non significativement probant.

Tel que présentée, cette relation est cependant insuffisante pour établir une comparaison de modèles ayant un nombre de paramètres différents. En effet, a coefficient concurrentiel égal, le principe du rasoir d'Occam nous suggère de ne retenir que le plus simple. Tel quel, la valeur du coefficient concurrentiel ne permet pas de tenir compte de ce principe. La recette classique est

2. Imposez dans l'Eq.2.10 $O_{M_1,M_2} = 3$ et vous obtenez $P(M_1|\boldsymbol{y}) = 0.75$

d'ajouter un terme *ad-hoc* de pénalisation, fonction du nombre de paramètres, ou d'utiliser un contexte Bayésien, ce dernier tenant compte naturellement du principe du rasoir d'Occam.

Qui plus est, l'approche MLE n'est fonctionnelle que lorsque la vraisemblance n'est pas une fonction multimodale. Comme nous le montrons Chapitre 4 sous certaines conditions, qui peuvent être liées soit à la qualité de la mesure soit aux propriétés physiques intrinsèques de l'astre étudié, ce n'est plus toujours le cas.

2.2 Approche Bayésienne simple : Approche MAP

L'approche Bayésienne la plus simple repose sur la recherche du maximum de la *probabilité a posteriori* (MAP pour *maximum a posteriori*) et non plus de la vraisemblance.

La principale caractéristique de cette approche consiste à être dans la mesure de tenir compte d'un savoir ou d'une connaissance *a priori* dans le calcul probabiliste. En se basant sur quelques règles issues de la logique et de la théorie des ensembles, l'approche Baysienne permet donc de réviser notre connaissance dite *a priori* (avant la confrontation avec les faits) pour en construire une nouvelle dite *a posteriori* (après la confrontation avec les faits).

2.2.1 Probabilité *a priori* et probabilité *a posteriori*

En théorie des ensembles l'union de deux sous ensembles A et B correspond à leur somme moins leurs parties communes. C'est la règle de l'addition en logique :

$$p(A \cup B) = p(A) + p(B) - p(A \cap B) \tag{2.11}$$

On peut aussi construire un nouvel ensemble sur la base de deux sous ensembles par adjonction de ceux ci : c'est la règle de multiplication en logique,

$$p(A \cap B) = p(A|B)p(B) = p(B|A)p(A) \tag{2.12}$$

Toute l'approche Bayésienne repose sur ces deux opérations logiques. De la première, on peut construire par exemple la règle,

$$\pi(A|B) + \pi(\bar{A}|B) = 1, \tag{2.13}$$

Où π représente une probabilité ou densité de probabilité. De la seconde, on peut établir la fameuse relation de Bayes,

$$\pi(A|B,C) = \frac{\pi(B|A,C)\pi(A|C)}{\pi(B|C)} = \frac{\psi(B,A|C)}{\int \pi(B|AC)\pi(A|C)dA} \tag{2.14}$$

$\pi(A|B,C)$ peut représenter par exemple, la probabilité (*a posteriori*) que l'événement A se réalise sachant qu'il existe un autre événement B [3]. A l'inverse, $\pi(B|A)$ représentera, la probabilité que l'événement B se réalise sachant qu'il existe A.

C pourra correspondre au contexte dans lequel s'établit ces opérations et constitue donc un *a priori*. Ce peut être par exemple, le modèle théorique choisi ou les instruments de mesures choisis.

Selon cette lecture de l'équation, $\pi(A)$ et $\pi(B)$ représentent les probabilités (*a priori*) respectives que chacun des événements A et B se réalisent, indépendamment les uns des autres.

Dans le jargon, on appelle *loi jointe*, le numérateur

$$\psi(B,A|C) = \pi(B|AC)\pi(A|C), \tag{2.15}$$

3. l'interprétation de C se fait quelques lignes plus loin.

égale au produit des *a priori* et de la vraisemblance. On appelle *loi marginale* le dénominateur

$$\pi(B|C) = \int \psi(B, A|C)dA = \int \pi(B|AC)\pi(A|C)dA. \tag{2.16}$$

Ce dernier terme est un facteur de normalisation.

Le contexte Bayésien est parfaitement adapté aux sciences car les connaissances scientifiques évoluent sans cesse. Les connaissances établies par un ensemble d'observateurs B peuvent par exemple servir de base à un autre observateur A, souhaitant confirmer ou infirmer les observations de B. Les observations de B sont alors un *a priori* pour A. L'*a priori* n'est donc pas forcement subjectif comme sous entendu dans le langage courant. Il peut avoir été formulé grâce à une mesure préalable, dont on souhaite tenir compte. Quelques distributions *a priori* possibles, utilisées dans le cadre de la thèse sont présentées en Annexe A.

Une présentation exhaustive des implications et applications possibles de l'approche Bayésienne peut être trouvée dans le cours de Master prodigué par Guy Demoment [4].

2.2.2 Application de l'approche Bayésienne au cas stellaire

Comment construire la loi de probabilité *a posteriori* dans le cas qui nous intéresse ?

Nous cherchons à déterminer quel est le meilleur jeu de paramètres $\boldsymbol{\theta}$ d'un modèle M en accord avec le spectre de puissance à notre disposition $\boldsymbol{y} = S(\boldsymbol{\nu})$.

On cherche donc à calculer la quantité $\pi(\boldsymbol{\theta}|\boldsymbol{y}, M, I)$. I sert à rappeler que tous les calculs s'effectuent dans un contexte donné, défini par un ensemble d'hypothèses (mesures, modèles,...), implicitement déclarées ou non. Appliquons la règle de Bayes (Eq. 2.14),

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}, M, I) = \frac{\pi(\boldsymbol{\theta}|M, I)\pi(\boldsymbol{y}|\boldsymbol{\theta}, M, I)}{\pi(\boldsymbol{y}|M, I)} = \frac{\psi(\boldsymbol{\theta}, \boldsymbol{y}|M, I)}{\pi(\boldsymbol{y}|M, I)} \tag{2.17}$$

On reconnait de suite le terme de vraisemblance $\pi(\boldsymbol{y}|\boldsymbol{\theta}, M, I)$ précédemment calculé, Chap.2.1, Eq.2.8. $\pi(\boldsymbol{\theta}|M, I)$ est l'*a priori* sur ce que vaut notre vecteur de paramètres $\boldsymbol{\theta}$. $\pi(\boldsymbol{y}|M, I)$ est souvent appelé *evidence* en anglais (la preuve en français) car il définit la probabilité que le modèle soit cohérent avec les observations, indépendamment des valeurs des paramètres. On le calcule en utilisant l'Eq.2.16.

Dans le cas de l'astérosismologie, les différents paramètres θ_i sont majoritairement indépendants les uns des autres. Ceci est vrai pour les fréquences individuelles des modes stellaire car le couplage entre modes est faible dans les étoiles de type solaire de la séquence principale (absence de modes mixtes mesurable par exemple). Mais ceci peut être discutable pour la hauteur et la largeur car, pour un mode donné ces deux paramètres sont corrélés. Dans ce cas, supposer l'indépendance de ces deux paramètres peut introduire des biais en fonction de la nature des *a priori* utilisés [5]. La règle du produit nous permet de formaliser cette hypothèse d'indépendance des variables *a priori*,

$$\pi(\boldsymbol{\theta}|M, I) = \prod_{i}^{i=N} \pi(\theta_i|M, I). \tag{2.18}$$

Tout comme l'approche maximum de vraisemblance (MLE), l'approche *Maximum a posteriori* (MAP) ne s'intéresse qu'au maximum du critère statistique. Dans ce cas, en tant que constante,

4. Disponible à l'adresse http ://djafari.free.fr/books/PolyProbasT1.pdf

5. Nous verrons par la suite que nous utilisons des *a priori* non-informatif pour la largeur et la hauteur pour éviter cette écueil.

l'*evidence* n'est pas évaluée. Sa formulation intégrale est de toute manière trop complexe pour être prise en compte par une approche MAP.

Finalement, nous voyons qu'entre l'approche fréquentiste MLE et l'approche Bayésienne MAP, le critère statistique ne diffère que de peu : le critère statistique utilisé est la *loi jointe* $\psi(\boldsymbol{\theta}, \boldsymbol{y}|M, I)$, défini comme le produit de la vraisemblance et des quantités *a priori*. La recherche du maximum se fera encore une fois, dans la grande majorité des cas, *via* des méthodes numériques.

2.2.3 Comparaison de modèles avec une approche MAP

De manière similaire à l'approche MLE, La comparaison de modèle pour une approche MAP repose sur la comparaison du critère statistique. Le coefficient concurrentiel n'est plus défini comme le rapport des maximum de vraisemblance mais comme celui des *lois jointes*,

$$O_{M_1,M_2} = \frac{\arg\max\ \{\psi(\boldsymbol{\theta_1}, \boldsymbol{y}|M_1, I)\}}{\arg\max\ \{\psi(\boldsymbol{\theta_2}, \boldsymbol{y}|M_2, I)\}}. \tag{2.19}$$

La relation 2.10 définissant la probabilité relative, reste valable.

Cette fois ci, le principe du rasoir d'Occam est implicitement implémenté dans cette approche. Pour comprendre comment, prenons un exemple simple.

Imaginons un problème dans lequel on cherche à définir si un signal est noyé dans un bruit blanc. On va supposer que le signal est caractérisé par un paramètre unique, une hauteur par exemple. Dans ce cas, le problème décisionnel se résume à comparer un premier modèle ne contenant que du bruit avec un autre contenant à la fois du bruit et le signal recherché. Si on suppose l'indépendance bruit/signal, avec une approche MAP, établir la présence de signal se résumera à calculer,

$$O_{M_{x+b},M_b} = \frac{\pi(\boldsymbol{s}|\hat{x}, \hat{b}, M_{x+b})\pi(\hat{b}|M_{x+b})\pi(\hat{x}|M_{x+b})}{\pi(\boldsymbol{s}|\hat{b}, M_b)\pi(\hat{b}|M_b)}. \tag{2.20}$$

Avec $\boldsymbol{s}$ les observations ($\boldsymbol{s} = \{s_0, ..., s_i, ..., s_N\}$), $\hat{x}$ et $\hat{b}$ les paramètres au maximum de l'*a posteriori*, respectivement associés au signal et au bruit.

On fera remarquer que l'indépendance des paramètres implique aussi que notre connaissance *a priori* sur le bruit doit être indépendante du signal recherché : $\pi(b|M_{x+b}) = \pi(b|M_b)$.

Pour faire simple, on va supposer que nos deux paramètres sont contraints par des *a priori* uniformes,

$$\begin{aligned} \pi(b|M_b) &= 1/(b_1 - b_0) \text{ si } b_0 \leq b \leq b_1 \\ &= 0 \text{ sinon} \qquad (2.21) \\ \pi(x|M_{x+b}) &= 1/(x_1 - x_0) \text{ si } x_0 \leq x \leq x_1 \\ &= 0 \text{ sinon} \qquad (2.22) \end{aligned}$$

On voit alors qu'à maximum de vraisemblance égal, le coefficient concurrentiel est égal à l'*a priori* $\pi(x|M_{x+b})$ et rend le rapport O_{M_{x+b},M_b} inférieur à 1.

En ce sens, l'approche par MAP tient naturellement compte de la complexité des modèles car les *a priori* jouent le rôle de *pénalisation*. Cependant, comme le niveau de pénalisation est étroitement lié aux *a priori* utilisés, les conséquences de leur choix doivent systématiquement être testées, surtout s'ils sont choisis de manière purement subjective. La plupart des auteurs recommandent d'évaluer l'influence de l'*a priori* tant sur les paramètres *a posteriori* estimés que sur la probabilité.

Le critère statistique utilisé (*loi jointe*) est fonctionnel si la distribution *a posteriori* dispose d'un unique maximum. Ceci est vrai uniquement si les *a priori* réduisent à l'unité les solutions possibles ou si la vraisemblance est déjà unimodale. Lorsque l'on ne dispose par d'une information préalable suffisamment forte, ces conditions peuvent ne pas être remplies. L'unique méthode fiable restante consiste à échantillonner la loi jointe $\pi(\boldsymbol{\theta}|\boldsymbol{y}, M, I)$. Durant la thèse, nous avons développé une telle approche, basée sur un échantillonnage mettant en oeuvre des chaines de Markov, car l'analyse des modes stellaires des étoiles CoRoT s'est révélée très délicate. Nous avons consacré les Chapitres 3 et 4 à ce type d'algorithme et à ces applications.

2.3 Le calcul d'erreur avec des algorithmes de type MLE ou MAP

Le calcul d'erreur, que ce soit pour un Bayésien ou un fréquentiste, repose sur notion de probabilité. Celle ci est un scalaire (P) obtenu par intégration sur un intervalle donné d'une fonction de densité de probabilité (pdf en anglais ou fpd en français) $\pi(x)$, $P = \int_{a_x}^{b_x} \pi(x)dx$. Si nous souhaitons estimer la valeur que prendra x avec un degré de confiance de $P\%$, ce qui compte ce n'est pas tant la moyenne (ou tout autre indicateur statistique comme la mediane, etc...) mais l'intervalle sur lequel l'intégrale est effectuée, nommé *intervalle de confiance*. La fonction densité de probabilité est normalisée par définition. Donc dans la limite où $a_x = -\infty$ et $b_x = +\infty$, on a $P = 100\%$.

Une interprétation du résultat de l'intégrale est : il est probable à $P\%$ que la valeur x prennent une valeur comprise entre a_x et b_x.

Cette notion d'intervalle de confiance est une notion inscrite dans les moeurs pour les Bayésiens. Elle existe aussi chez les fréquentistes, mais son interprétation est plus limitée.

En effet, nous pouvons définir deux types de probabilité :

- la probabilité vue comme la proportion d'intervalles qui contiennent le paramètre pour un grand nombre de mesures : elle est appelée *probabilité de recouvrement fréquentiste*. Cette définition n'est pas exclue dans le contexte Bayésien, mais n'est pas intuitive pour le scientifique.
- la probabilité (Bayésienne) que l'intervalle contienne la valeur correcte du paramètre. C'est l'interprétation courante qu'un scientifique fait de la probabilité.

Une conséquence directe est qu'il n'est pas conceptuellement problématique d'attribuer une probabilité sur la base d'une unique mesure pour un Bayésien alors que ça l'est pour un fréquentiste.

Quoi qu'il en soit, dans le cas particulier où les distributions sont gaussiennes, l'approche "classique" pour estimer l'intervalle de confiance consiste à définir l'erreur de mesure comme l'écart-type de la distribution, en centrant l'intervalle sur la valeur moyenne. Autrement dit, on cherche les bornes telle que : $P = \int_{\mu-\sigma}^{\mu+\sigma} \pi(x)dx = 66.7\%$. μ représente la valeur la plus probable qui est aussi par définition la position du maximum et de la médiane de la pdf pour une distribution gaussienne. Si la valeur la plus probable n'est pas confondue avec la moyenne alors l'estimateur est biaisé.

C'est ce qui peut arriver dès lors que les distributions ne sont plus gaussiennes. Pour les distributions unimodales, la plupart du temps un changement de variable du type $x \longrightarrow ln(x)$ ou $x \longrightarrow 1/x^n$, permet d'approximer les distributions par des distributions gaussiennes. Dans de tels cas, on peut définir des barres d'erreurs asymétriques et la vraisemblance/probabilité n'est plus un estimateur biaisé. Par contre ce n'est plus le cas lorsque l'on n'est pas capable de se ramener à des lois normales ; par exemple avec des fonctions multimodales que l'on rencontrera lors de l'analyse des signaux stellaires. Une estimation fiable de l'erreur sur le paramètre ne peut

alors qu'être obtenue par échantillonnage, méthode statistique bien plus couteuse en temps de calcul que les méthode par recherche de maximum du type MLE ou MAP.

Pour ces dernières, le calcul de l'erreur repose sur le calcul de la matrice Hessienne du critère statistique $L(x)$. Pour comprendre quelle en est la raison et les hypothèses sous-jacentes, nous allons tout d'abord nous placer dans le cas simple d'une fonction vraisemblance univariée $L(x)$, de paramètre unique x (scalaire).

Si on émet l'hypothèse qu'au voisinage du maximum de vraisemblance ou de probabilité (noté $L(x_0)$), l'espace peut être localement décrit par une gaussienne, alors l'opposé du log-vraisemblance est localement une parabole,

$$-ln(L(x)_{\sim L(x_0)}) = \frac{1}{2}\frac{(x-x_0)^2}{\sigma^2}. \tag{2.23}$$

On remarque alors que la dérivée seconde de $-ln(L(x)_{\sim L(x_0)})$ n'est rien d'autre que l'inverse de la variance,

$$-\frac{d^2 ln(L(x)_{\sim L(x_0)})}{dx^2} = \frac{1}{\sigma^2}. \tag{2.24}$$

Dans le cas d'un espace de paramètre de dimension $dim(x) > 1$ on a,

$$-ln(L(\boldsymbol{x})_{\sim L(\boldsymbol{x_0})}) = \frac{1}{2}\left\langle \boldsymbol{x}-\boldsymbol{x_0}\right| \Sigma^{-1} \left|\boldsymbol{x}-\boldsymbol{x_0}\right\rangle, \tag{2.25}$$

où Σ est la matrice de *variance-covariance* (ou de *covariance*). Elle est ainsi appelée car à sa diagonale correspondent les termes de variance de σ_{ii} ; des éléments non diagonaux correspondent aux termes de covariance $cov(x_i, x_j)$. Cette matrice est définie positive symétrique ($cov(x_i, x_j) = cov(x_j, x_i)$).

Σ^{-1} est souvent nommé matrice de covariance à tort puisque c'est plutôt l'*inverse de la matrice des covariances* (ou encore la *la comatrice des covariances*). Par extension et en s'inspirant de l'équation 2.24, on se rend compte que la Hessienne s'exprime,

$$H = -\frac{\partial^2 ln(L(\boldsymbol{x}))}{\partial \boldsymbol{x_i} \partial \boldsymbol{x_j}} = \Sigma^{-1}. \tag{2.26}$$

Autrement dit, si le critère statistique utilisé est localement gaussien au voisinage de la solution $\boldsymbol{x_0}$, alors l'inverse de la Hessienne est égale à la matrice des covariances. En générale les erreurs à 1σ fournies par le calcul de la Hessienne sont sous-estimées : l'hypothèse de gaussianité n'est souvent pas pleinement justifiée, ce qui milite en faveur des méthodes d'échantillonnage statistique.

En fait, il existe un critère, appelé *Borne de Cramer-Rao* ou *inégalité de Cramer-Rao* (Aitken & Silverstone 1942; Cramer 1946) définissant la limite théorique atteignable par la variance, connaissant un critère statistique sans biais. Son calcul se base sur la *quantité d'information de Fisher*, grandeur introduite par le statisticien du même nom et permettant une quantification de l'information contenue dans une distribution statistique. Cette notion a été introduite parallèlement à la notion de maximum de vraisemblance au début du 20ème siècle avec une approche fréquentiste, mais ces conclusions sont tout aussi valides pour un Bayésien.

On défini la *quantité d'information de Fisher* (ou *matrice de Fisher*) $\mathcal{I}(\boldsymbol{\theta})$ par l'espérance du produit croisé des dérivées partielles de la loi de densité de probabilité $p(\boldsymbol{x}, \boldsymbol{\theta})$

$$\mathcal{I}(\theta_i, \theta_j) = \mathbb{E}\left[\left(\frac{\partial}{\partial \theta_i} ln\, p(\boldsymbol{x}, \boldsymbol{\theta})\right)\left(\frac{\partial}{\partial \theta_j} ln\, p(\boldsymbol{x}, \boldsymbol{\theta})\right)\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} ln\, p(\boldsymbol{x}, \boldsymbol{\theta})\right] \tag{2.27}$$

où $\boldsymbol{x}$ et $\boldsymbol{\theta}$ représentent respectivement la variable aléatoire et les paramètres (d'un modèle non explicité ici) pour la décrire.

L'inégalité de Cramer-Rao se formule,

$$Var[\boldsymbol{\theta}] \geq \frac{1}{\mathcal{I}(\boldsymbol{\theta})}. \tag{2.28}$$

Si la loi de probabilité p suivi par x est une loi normale multivariée de matrice de covariance $\boldsymbol{\Sigma}$ et si l'estimateur est non biaisé alors,

$$Var[\boldsymbol{\theta}] = \boldsymbol{\Sigma} = \frac{1}{\mathcal{I}(\boldsymbol{\theta})}. \tag{2.29}$$

En somme, la précision d'une détermination d'un ensemble de paramètre θ décrivant une variable aléatoire $\boldsymbol{x}$ est limité par l'inverse de la matrice de covariance, qui n'est rien d'autre que la Hessienne. **De ce point de vue, on remarquera encore une fois que la Hessienne ne fournit qu'une limite inférieure des barres d'erreurs**. Une description plus détaillée de ces aspects statistiques[6] peut être trouvée dans Kendall & Stuart (1967).

6. On trouvera aussi des notes de cours intéressantes sur le sujet à l'addresse http ://www.ensta.fr/ pcarpent/MNOS/PDF/12-Cramer.pdf

Chapitre 3

L'échantillonnage par Markov Chain Monte-Carlo

Sommaire

3.1 Introduction sur les méthodes d'échantillonnage 41

3.1.1 Echantillonnage par réjection 42

3.1.2 Qu'est ce qu'un processus Markovien ? 43

3.2 L'Echantillonage de Monte-Carlo par Chaine de Markov 44

3.2.1 Définitions et propriétés : des chaines de Markov à l'algorithme de Metropolis-Hasting 44

3.2.2 Quel(s) critère(s) pour assurer une exploration optimale de l'espace des paramètres ? 47

3.2.3 L'Adaptive Markov Chain Monte-Carlo : un moyen de contrôle automatique du taux d'acceptation/réjection 54

3.2.4 Eviter le confinement dans des maxima locaux du critère statistique . . 61

3.2.5 Comparaison de modèles avec un algorithme MCMC tempéré 64

3.2.6 Résumé sur la méthode d'échantillonnage 68

3.1 Introduction sur les méthodes d'échantillonnage

Lors de la présentation des méthodes "classiques" de recherche d'optimum statistique, nous avons montré qu'il était assez simple de déterminer le meilleur ajustement des données à condition que le critère statistique (vraisemblance ou probabilité *a posteriori*) ne soit pas multimodal. La comparaison de modèles dans les approches par maximum de vraisemblance nécessite que les tailles des espaces des paramètres soient identiques. On peut prendre pour exemple le cas d'ajustement polynomial sur une fonction $f(x)$: il est toujours possible de trouver un polynôme de degré p qui décrit **exactement** tous les points de mesures x_i. Mais il n'est pas crédible de comparer directement [1] les vraisemblances de deux modèles de complexités différentes avec une approche par maximum de vraisemblance (par exemple les χ^2 obtenues avec deux ajustements polynomiales de degrés différents ne sont pas comparables) . Dans les approches par maximum de probabilité *a posteriori*, cette condition n'est plus requise, grâce à une quantification des

1. Il existe tout de même des méthodes ad-hoc, *a posteriori*, pour choisir quel niveau de complexité du modèle est suffisant pour décrire des observations avec une approche fréquentiste.

connaissances a priori sur le jeu de paramètres utilisés pour décrire les modèles qui incorpore implicitement le principe du rasoir d'Occam [2].

Cependant, bien que permettant d'établir de manière objective une comparaison de modèles (à partir d'information *a priori* pouvant être subjective), les approches MAP tout comme MLE nécessitent une connaissance *a priori* relativement bonne sur les paramètres. En effet, sur quels critères devons nous définir la position initiale dans l'espace des paramètres, lorsque nous ne disposons que d'une connaissance vague, voire nulle, des paramètres recherchés ? Même des approches par maximum de probabilité peuvent être mises en déroute dans ce cas. L'absence d'information *a priori* contraignante augmente le volume d'espace à explorer, qui peut éventuellement contenir de multiples maxima.

Dans ces conditions, il ne faut plus se limiter au seul maximum du critère statistique. La structure globale de celui ci doit être intégralement déterminé. Toutefois, à de très rares exceptions près, il est impossible de calculer analytiquement la vraisemblance et encore moins la probabilité *a posteriori*. Obtenir la structure globale passe donc nécessairement par l'échantillonnage du critère statistique.

3.1.1 Echantillonnage par réjection

De multiples méthodes d'échantillonnage existent. La plus connue d'entre elles, appelée *échantillonnage par réjection* ou *Monte-Carlo*, consiste à échantillonner une fonction $f(x)$ en tirant aléatoirement des nombres dans un espace délimité par une seconde fonction $q(x)$ bornée. On peut comparer cette méthode à un lancé de fléchettes sur une cible. Les fléchettes tombant sur la cible adhèrent : ce sont des échantillons retenues. Celles se heurtant au mur, support de la cible n'adhèrent pas et tombent au sol : ce sont des échantillons rejetés. Plus précisément, supposons que nous souhaitions échantillonner une fonction $f(x)$ connu à une constante près, à partir d'une autre fonction, de densité $q(x)$. Supposons que l'on s'arrange pour que la fonction d'échantillonnage $q(x)$ multipliée par une constante finie A, soit une fonction enveloppe de $f(x)$ [3]. Le volume dans lequel s'effectue l'échantillonnage est alors délimité par $Aq(x)$ et les bornes a, b de $q(x)$ (cf. Fig.3.1).

Dans le cas le plus simple $q(x)$ est une fonction de distribution uniforme sur l'intervalle $x \in [a, b]$: $q(x) = 1/(b - a)$. L'algorithme se résume en 3 étapes :

1/ On tire aléatoirement x_i selon la distribution $x_i \sim q(x)$

2/ On tire aléatoirement u_i selon une loi uniforme $u_i \sim U[0, 1]$.

3/ On accepte x_i si $u_i < f(x_i)/A$, autrement on la rejette.

Une des applications courantes de la méthode par réjection est de fournir une évaluation d'une fonction intégrale lorsqu'une approche analytique est impossible. Ce type d'algorithme est efficace lorsque le volume à échantillonner est plutôt restreint, mais devient très vite gourmand en temps de calcul pour des espaces de paramètres importants. De plus, il n'est pas toujours possible de borner la fonction à échantillonner avec une valeur de A raisonnablement proche du maximum m de $f(x)$: lorsque $A >> m$ la probabilité d'accepter un nouvel échantillon (étape 3) peut devenir ridiculement petite (proportionnelle aux rapport des volumes d'espaces $\int Aq(x)dx$ et $\int f(x)dx$).

Il existe plusieurs alternatives pour échantillonner une fonction, mais les procédures basées sur des chaines de Markov font probablement partie des méthodes les plus efficaces et universelles car

2. Pour rappel, le principe du rasoir d'Occam, stipule qu'il faut toujours choisir la solution la plus simple parmi un ensemble de solutions possible. En somme, un modèle relativement simple reproduisant des observations est toujours préférable à un modèle plus complexe, reproduisant avec une qualité identique les mêmes observations.

3. En d'autre terme il faut que $f(x) \leq Aq(x)$

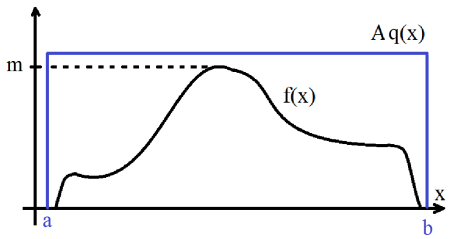


FIGURE 3.1 – Illustration du principe d'échantillonnage de Monte-Carlo. $f(x)$ est la fonction à échantillonner, $q(x)$ la fonction de densité servant à échantillonner, et A une constante assurant que $Aq(x)$ est une fonction enveloppe de $f(x)$. Idéalement il faut $max\{Aq(x)\} = max\{f(x)\} = m$.

elles ne nécessitent théoriquement aucune connaissance préalable sur la fonction à échantillonner.

3.1.2 Qu'est ce qu'un processus Markovien ?

Les méthodes Monte-Carlo ont initialement été développées à Los Alamos durant la seconde guerre mondiale, dans le cadre du projet Manhatan par Stanislaw Ulam et John von Neumann. La méthode par réjection présentée précédemment fut initialement appliquée pour le calcul des taux de multiplication des neutrons lors des réactions de fission. Quelques années après, une approche alternative plus efficace (car ne requérant pas une connaissance préalable d'une borne A de la fonction à échantillonner) a été proposée par Nicholas Metropolis (Metropolis et al. 1953), toujours dans le cadre de projets militaires associés au développement des bombes atomiques et plus généralement, à la physique atomique. Nicholas Metropolis cherchait à calculer les propriétés à l'équilibre thermodynamique d'un système physique. Il a proposé que le calcul devait se faire en explorant l'espace des paramètres de manière à minimiser les variations d'énergie du système ΔE. Il a alors suggéré que cette recherche se fasse en acceptant un nouvelle position dans l'espace des paramètres avec une probabilité $e^{-\Delta E/T}$. T correspond ici à la température d'équilibre du système. Cette idée est maintenant la base de beaucoup de procédures d'échantillonnage par chaine de Markov. En effet, ce processus de recherche du minimum d'énergie est Markovien, au sens ou la position dans l'espace paramètre à l'itération i, ne dépend que de la position à l'itération $i-1$. On parle parfois de processus à mémoire faible. La transition entre des positions successives est probabiliste et forme un processus aléatoire. Un autre exemple physique de processus Markovien est la diffusion d'une particule dans un milieu. En première approximation, un processus de diffusion est constitué d'une successions de collisions. Chaque collision efface l'information sur la trajectoire de la particule. D'ailleurs, les approches théoriques permettant de définir le régime de fonctionnement idéal (*i.e.*, régime dans lequel l'échantillonnage est le plus efficace) d'une chaine de Markov repose pour beaucoup sur l'analogie avec les processus de diffusion (cf. Chap.3.2.2).

Le développement des méthodes Monte-Carlo coïncide avec les premières heures du développement des ordinateurs (ENIAC en 1947 puis ensuite MANIAC en 1952). Cependant les MCMC n'ont eu que peu d'échos dans la communauté scientifique durant de longues années d'une part, car la puissance de ces machines était dérisoires (5 heures avec ENIAC pour effectuer une simulation Monte-Carlo avec quelques dizaines d'échantillons !) ; d'autre part, car les recherches théoriques sur les systèmes Markovien étaient balbutiantes. Ce n'est qu'après les travaux de Hasting dans les années 1970 (qui a généralisé et mieux formalisé l'algorithme de Metropolis)

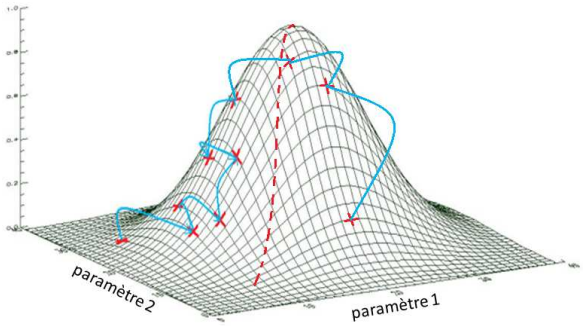


FIGURE 3.2 – Exemple de critère statistique pour un espace des paramètres 2D. Dans le cas d'un échantillonnage par un processus Markovien (croix rouges et arcs bleus), l'exploration de l'espace des paramètres s'effectue par sauts successifs. Les coordonnées de la position à l'itération i dépendent exclusivement de la position à $i-1$. Lors de la recherche du maximum (pointillés rouges), le processus est purement Hamiltonien.

que le sujet pris de l'importance dans les communautés scientifiques. Pour le lecteur souhaitant approfondir les aspects historiques sur le développement des chaines de Markov, on citera l'article de Roger Eckhardt (Eckhardt 1987) ou encore Robert & Casella (2008).

On ajoutera que les systèmes Markoviens peuvent être opposés aux processus Hamiltoniens, parfaitement déterministes. Au cours de leur évolutions de tels systèmes entretiennent une mémoire forte de leurs états antérieurs. On peut citer le très classique exemple des trajectoires balistiques. La grande majorité des procédures basées sur la recherche du maximum (MLE ou MAP par exemple) résolvent un système d'équations différentielles (recherche des zéros des dérivées partielles du critère statistique, cf. Chap.2.1) par des méthodes numériques du type descente de gradient. Ce sont donc des processus purement Hamiltoniens.

Ces dernières années, avec l'essor des approches Bayésiennes, l'échantillonnage de Monte-Carlo par chaine de Markov (MCMC pour Markov Chain Monte-Carlo en anglais) a conquis la plupart des disciplines scientifiques. A partir du moment où le problème se résume à une sélection entre modèles (qui consiste en un calcul d'intégrale comme nous le verrons plus loin) ou à l'évaluation de paramètres dans des situations complexes, l'échantillonnage par MCMC est l'une des solutions les plus efficaces de calcul. Il existe dans la pratique plusieurs implémentations possibles des MCMC, qui se différencient par la stratégie adoptée pour résoudre les problèmes inhérents à l'analyse de fonction par MCMC. Dans ce qui suit, nous présenterons les propriétés fondamentales des chaines de Markov. En partant de l'algorithme de Metropolis, l'un des plus simple qui soit, nous aborderons les difficultés techniques propres au MCMC, puis nous présenterons les solutions algorithmiques que nous avons adoptées pour les résoudre.

3.2 L'Echantillonage de Monte-Carlo par Chaine de Markov

3.2.1 Définitions et propriétés : des chaines de Markov à l'algorithme de Metropolis-Hasting

Cette section vise à introduire le vocabulaire associé aux chaines de Markov puis tente d'expliquer comment l'algorithme de Metropolis-Hasting est formulé d'un point de vue statistique. Cette partie est loin d'être exhaustive. Au contraire, elle ne fait que survoler les fondamentaux

sur les chaines de Markov. Au lecteur souhaitant approfondir les aspects théoriques sous-jacents aux MCMC et plus généralement aux chaines de Markov, nous recommandons de lire le livre de Gamerman (1997).

Comme déjà souligné, une chaine de Markov est un processus stochastique (caractérisé par une *probabilité de transition* P_r) pour lequel les états, passés, présent et futures sont indépendants. Cette phrase peut être formellement retranscrite par la relation,

$$P_r(\theta^{(n+1)} = y|\theta^{(n)} = x,\, \theta^{(n-1)}, ...,\, \theta^{(0)}) = P_r(\theta^{(n+1)} = y|\theta^{(n)} = x). \tag{3.1}$$

Les valeurs de θ sont définies dans un ensemble S (l'ensemble des valeurs possibles) d'une distribution de probabilité P : $\theta_0, ..., \theta_{n-1}, \theta_{n+1} \subset S$.

On défini alors un *noyau de transition* (*transition kernel*) comme étant la fonction de transition d'un état x à un état y, $P(y|x)$ qui satisfait les conditions suivantes,

- $P(y|x)$ est positif : $\forall x, y \in S, P(y|x) \geq 0$.
- $P(y|x)$ est une fonction normalisée : $\forall x \in S, \int_{y \in S} P(y|x) = 1$.

Moins formellement, ce noyau de transition correspond à l'ensemble des transitions que l'on autorise dans l'espace des paramètres à explorer. Pour mieux comprendre prenons pour exemple un problème où les états possibles sont discrets et dénombrables : le *processus de naissance et mort* (*birth-death process*). Ce type de processus Markovien discret a énormément d'applications (biologie/génétique, démographie, etc...). A chaque instant, on peut construire un graphe des transitions possibles. 4 transitions sont envisagées, dont 3 sont autorisées ici. A chacune de ces transitions sont affectés des probabilités d'occurrences de la transition.

- On a une naissance si le nouvel état y de la chaine répond à la relation $y = x + 1$. Cette transition arrive avec une probabilité $P(y|x) = p_1$.
- On a un décès si y = x -1. Cette transition arrive avec une probabilité $P(y|x) = p_2$.
- L'état précédent n'est pas modifié : y = x. Cette transition arrive avec une probabilité $P(y|x) = p_3$.
- Tout autre transition est interdite : $P(y|x) = 0$.

On a bien entendu, $p_1 + p_2 + p_3 = 1$. Dans les cas discret il est courant d'utiliser une représentation matricielle pour décrire l'ensemble des transitions permises. La probabilité de passer d'un état x à un état y, partant de l'itération n jusqu'à l'itération m est définie par la relation de Chapman-Kolmogorov elle même directement déduite des matrices de transitions $P^n(x|z)$, $P^m(z|y)$,

$$P^{n+m}(y|x) = \sum_z P^n(y|z)P^m(z|x). \tag{3.2}$$

On fera remarquer que ces matrices contiennent des valeurs fluctuantes en fonction des itérations : ce sont des matrices stochastiques : chaque matrice est un instantané des états visités par la chaine de Markov. La relation de Chapman-Kolmogorov est fondamentale pour expliquer les propriétés de convergence des matrices stochastique de transition.

Dans le cas de variables continues (cas qui nous intéresse tout particulièrement), il n'est pas possible de construire de tels matrices de transition et on définit la distribution conditionnelle,

$$\forall x, y \in S,\, \exists\, p(y|x) = \frac{\partial P(y|x)}{\partial y}. \tag{3.3}$$

L'equation de Chapman-Kolmogorov s'écrit cette fois sous forme intégrale,

$$P^{n+m}(y|x) = \int_{-\infty}^{+\infty} p^n(z|x)P^m(y|z)dz. \tag{3.4}$$

Notion de distribution stationnaire. Un problème fondamental des chaines de Markov consiste à établir leurs propriétés asymptotiques lorsque le nombre d'itération $N \to \infty$. Autrement dit, connaissant un schéma/noyau de transition $p(y|x)$ (par exemple défini comme dans le processus de naissance/décès), existe il un état d'équilibre de la chaine ? si oui quel est asymptotiquement l'état final atteint ? Ceci introduit la notion de *distribution stationnaire* : sous certaines conditions, une distribution d'équilibre existe. On dit alors que la distribution stationnaire est atteinte lorsque le noyau de transition n'affecte plus, en moyenne, l'état d'équilibre π de la chaine de Markov,

$$\int_{-\infty}^{+\infty} \pi(x)p(y|x)dx = \pi(y) \text{ (dans le cas continu).} \tag{3.5}$$

Réversibilité ou "detailed balance". On dit qu'une chaine de Markov est réversible si elle remplit la *condition de réversibilité* :

$$\pi(x)p(y|x) = \pi(y)p(x|y). \tag{3.6}$$

La réversibilité assure l'existence d'une distribution stationnaire π.

Irréductibilité. La principale condition nécessaire pour qu'une chaine de Markov converge vers une distribution stationnaire, est l'*irréductibilité*. Une chaine de Markov est irréductible si il est permis d'atteindre à partir d'un état quelconque, un autre état quelconque, durant un nombre fini d'itérations et selon des probabilités strictement positives.

Récurrence. La propriété de *récurrence* est remplie si, les trajectoires passent une infinité de fois dans un ensemble de probabilité non nulle. Autrement dit, il est permis (et nécessaire) de visiter un état de manière répété.

Apériodicité. Une chaine de Markov est *apériodique* si aucun noyau de transition n'induit de comportement périodique des trajectoires. Par exemple une suite d'échantillons $\{x_0, x_1, ..., x_i\}$ ne doit pas se répéter dans un ensemble de N échantillons.

Ergodicité. Une chaine de Markov est dite *ergodique* si la chaine est à la fois *irréductible*, *apériodique* et *récurrente*. Autrement dit, il faut que la chaine de Markov puissent visiter tous les états possibles de la distribution d'équilibre de manière répétée mais non périodique. Si tel est le cas, il est alors garanti que la chaine possède une seule loi stationnaire. Une conséquence fondamentale de cette propriété est que les échantillons d'une chaine de Markov sont représentatifs de la loi stationnaire (π) au bout d'un nombre infini d'échantillons avec une probabilité de 1.

On parle de Monte-Carlo par Chaine de Markov lorsque une chaîne de Markov est ergodique car c'est à cette seule condition qu'elle est représentative de la distribution stationnaire. L'ergodicité est donc une propriété importante. Dans ce contexte la distribution stationnaire étant aussi la distribution que nous cherchons à échantillonner, on l'appellera *distribution d'intérêt*.

Maintenant que la notion de MCMC a été introduite tant du point de vue historique que statistique, nous allons présenter l'algorithme de Metropolis-Hasting.

Celui ci est réversible et son noyau de transition $p(y|x)$ est défini par 2 termes : une *distribution de proposition* $q(x, y)$ et une *probabilité de transition* $\alpha(x, y)$,

$$p(y|x) = q(y|x)\alpha(y|x). \tag{3.7}$$

W. Keith Hasting (Hastings 1970) a proposé que la probabilité d'accepter une transition $\alpha(y|x)$ soit de la forme,

$$\alpha(y|x) = min\left\{1, \frac{\pi(y)q(y|x)}{\pi(x)q(x|y)}\right\}, \tag{3.8}$$

de tel sorte que combinée à n'importe quelle distribution de proposition $q(x, y)$, la chaine est réversible. La chaine de Markov définie comme tel est aussi ergodique. Dans le cas particulier où la loi de proposition est choisie symétrique, c'est à dire que $q(y|x) = q(x|y)$ (c'est le cas des distributions gaussiennes), alors la chaine de Markov correspond à celle originellement élaborée par Metropolis et al. (1953). Dans ce cas, la probabilité de transition est simplement,

$$\alpha(y|x) = min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}. \tag{3.9}$$

Dans la pratique l'algorithme de Metropolis-Hasting est relativement simple à implémenter :

0. Initialisation de x : $x_{(0)} \leftarrow x_0$ et $i = i + 1$
1. Tirage d'un nombre aléatoire selon la loi de proposition : $r \sim q(y|x)$.
2. On propose une nouvelle position selon le schéma : $y = x_{(i-1)} + r$.
3. Evaluation de $\alpha(y|x) = \frac{\pi(y)q(y|x)}{\pi(x)q(x|y)}$.
4. Tirage d'un nombre aléatoire selon une distribution uniforme : $p \sim U[0, 1]$.
5. Si $p \leq \alpha(y|x)$ alors $x_{(i)} \leftarrow y$, Sinon $x_{(i)} \leftarrow x_{(i-1)}$
6. Tant que $i \leq N$ faire $i = i + 1$ et retourner à l'étape 1. Sinon, fin du programme.

Ce schéma numérique peut facilement être interprété graphiquement. Contrairement à un processus Hamiltonien du type recherche de la plus grande pente (*gradient descent*), d'itération en itération on ne converge pas forcément vers le maximum mais on s'autorise plus de liberté : il est possible d'aller en contre-sens du gradient (cf. Fig.3.2). Plus précisément, si la nouvelle position y correspond à une valeur $\pi(y)q(y|x) > \pi(x)q(x|y)$, on accepte toujours y. Sinon, on ne le fait qu'avec une probabilité proportionnelle au rapport des "hauteurs" $\pi(y)/\pi(x)$.

La principale difficulté de l'algorithme de Metropolis-Hasting réside dans le choix de la loi de proposition $q(x, y)$. La plupart du temps, celle ci est choisie gaussienne car d'une part cela simplifie l'algorithme et d'autre part cette loi est facilement paramétrisable. Cependant, comme nous le verrons au chapitre suivant, ce choix n'est pas aussi évident qu'il n'y parait et peut être inadéquat voir catastrophique pour la qualité de l'échantillonnage : en effet, ergodicité ne signifie certainement pas efficacité.

3.2.2 Quel(s) critère(s) pour assurer une exploration optimale de l'espace des paramètres ?

Comme nous l'avons suggéré, la progression d'une chaine de Markov dans l'espace des paramètres est entièrement définie par la *loi de proposition* choisie. C'est en fait elle qui définit la probabilité de transition globale de la chaine de Markov et contrôle l'efficacité de l'échantillonnage. Dans ce qui suit, nous présenterons comment cette efficacité est affectée par la loi de proposition.

Pour comprendre la nature des problèmes liés au choix de la loi de proposition, nous allons nous servir d'un exemple simple. Tout d'abord, construisons des données artificielles issues d'un

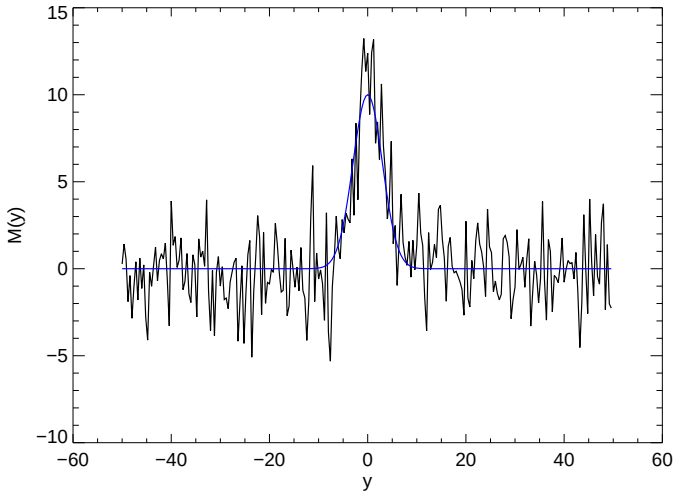


FIGURE 3.3 – Les données (en noir) et le modèle (en bleu) utilisé pour évaluer en première approche l'influence de la loi de proposition sur l'échantillonnage.

instrument fictif quelconque. Le modèle représentant les données est une fonction gaussienne, $M(x, H, \sigma, \mu)$, caractérisé par sa hauteur H, sa largeur σ et sa position centrale μ. x correspond à l'axe des abscisses.

On simule les perturbations dues au bruit instrumental en perturbant le profil idéal $M(x, H, \sigma, \mu)$ avec une variable aléatoire dont la distribution suit une loi normale,

$$y_i = M(x_i, H, \sigma, \mu) + N(0, \sigma_\epsilon^2) \tag{3.10}$$

On obtient ainsi des données simulées y dont l'allure est représentée Fig.3.3. Chaque paramètre va être caractérisé par une distribution des erreurs gaussienne dont la valeur centrale est la valeur la plus probable de la variable. Nous cherchons à établir quelle est la distribution de densité de probabilité de chacune des variables. On doit donc construire un critère statistique dont l'expression dépend des caractéristiques associées au bruit. Pour simplifier le problème, on fixe la hauteur durant l'échantillonnage. Il n'y a ainsi que deux variables, σ et μ, dont les écarts types seront notés par la suite ϕ_σ et ϕ_μ. Le critère statistique est construit de la même manière qu'au Chap.2.1. Dans le cas présent, la fonction que l'on cherche à échantillonner n'est rien d'autre que le critère du χ^2 (ne pas confondre avec la statistique du même nom),

$$l(y|H, \sigma, \mu) = -\sum_{i=0}^{N-1} (y_i - M(x_i, H, \sigma, \mu))^2 \tag{3.11}$$

Pour les besoins de notre démonstration, on se place dans le cas où la loi de proposition $q(\boldsymbol{\theta})$ est une loi normale multivariée de matrice de covariance diagonale. Ce cas est celui qui a été le plus étudié. La transition entre deux états sera intégralement contrôlé par la matrice de covariance. On notera σ_q^σ et σ_q^μ les largeurs caractéristiques de la loi de proposition pour chacune des variables (*i.e.* les deux termes diagonaux).

On observe alors le comportement des 2 variables au cours de l'échantillonnage (20 000 itérations), dans 3 situations différentes :

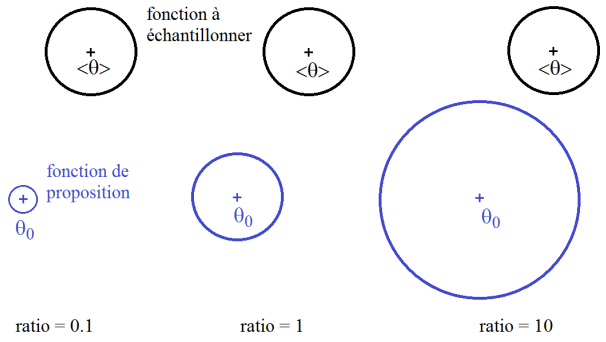


FIGURE 3.4 – Représentation imagée des trois conditions envisagés pour évaluer l'influence de la variance de la fonction de proposition normale. Le ratio se réfère au rapport entre les écarts types de la loi de proposition et de la fonction à échantillonner.

- Les écarts types de la loi de proposition (σ_q^σ et σ_q^μ) sont pris égaux à $1/10^{ème}$ des écarts types respectifs de chacune des variables (ϕ_σ et ϕ_μ).
- Les écarts types de la loi de proposition sont égaux aux écarts types respectifs de chacune des variables.
- Les écarts types de la loi de proposition sont 10 fois supérieurs aux écarts types respectifs de chacune des variables.

La figure 3.4 est une représentation graphique des trois situations envisagées. La figure 3.5 représente la manière dont a évolué la chaine de Markov dans l'espace des paramètres 2D constitué des variables σ et μ, pour chacune des situations. Quelque soit le rapport des écarts types entre loi de proposition et loi échantillonnée, on constate qu'il y a deux phases. Dans un premier temps la chaine progresse de manière quasi-linéaire vers la position du maximum du critère statistique. Cette phase est nommé *Burning phase* en anglais (temps de chauffage). La chaine de Markov reste ensuite confinée au voisinage du maximum. C'est la phase d'échantillonnage proprement dit. En observant plus attentivement le comportement de chacune des situations, on remarque que le cas où le rapport est de $1/10^{ème}$, la chaine prend beaucoup de temps à converger vers le maximum et l'échantillonnage une fois au maximum, n'est pas homogène : l'effet mémoire est relativement fort et la corrélation entre la $i^{ème}$ itération et les précédentes est importante. Dans le cas où le rapport est de 1, la chaine converge rapidement et l'échantillonnage est très homogène. Enfin si le rapport est de 10, le nombre de transition est faible et la chaine reste longtemps à une position donnée de l'espace des paramètres. Ce phénomène est plus net sur la Fig.3.6. L'échantillonnage semble tout de même plus homogène que dans le cas $1/10^{ème}$.

On en déduit que lorsque la fonction de proposition est une loi normale et que l'on échantillonne une distribution de la même forme bivariée, il existe un régime dans lequel l'échantillonnage doit être optimum. Cette exemple suscite plusieurs questions. Existe il un critère synthétique décrivant l'efficacité de la chaine de Markov ? Qu'en est-il lorsque l'on augmente la dimensionnalité du problème ou lorsque l'on change la forme de la distribution à échantillonner ?

Les premiers à avoir présenté des réponses concrètes à ces questions furent G.O. Roberts (Roberts et al. 1994, 1997) et A. Gelman (Gelman et al. 1996). Ils ont montré qu'un processus markovien basé sur l'algorithme de Metropolis est analogue à une diffusion de Langevin, décrivant le mouvement brownien d'une particule. Ils en ont déduit que lorsque la fonction statistique à échantillonner est une gaussienne multivariée sphérique (variance de chaque paramètre égales) de dimension d, l'efficacité de l'échantillonnage est contrôlée par la vitesse de diffusion dans

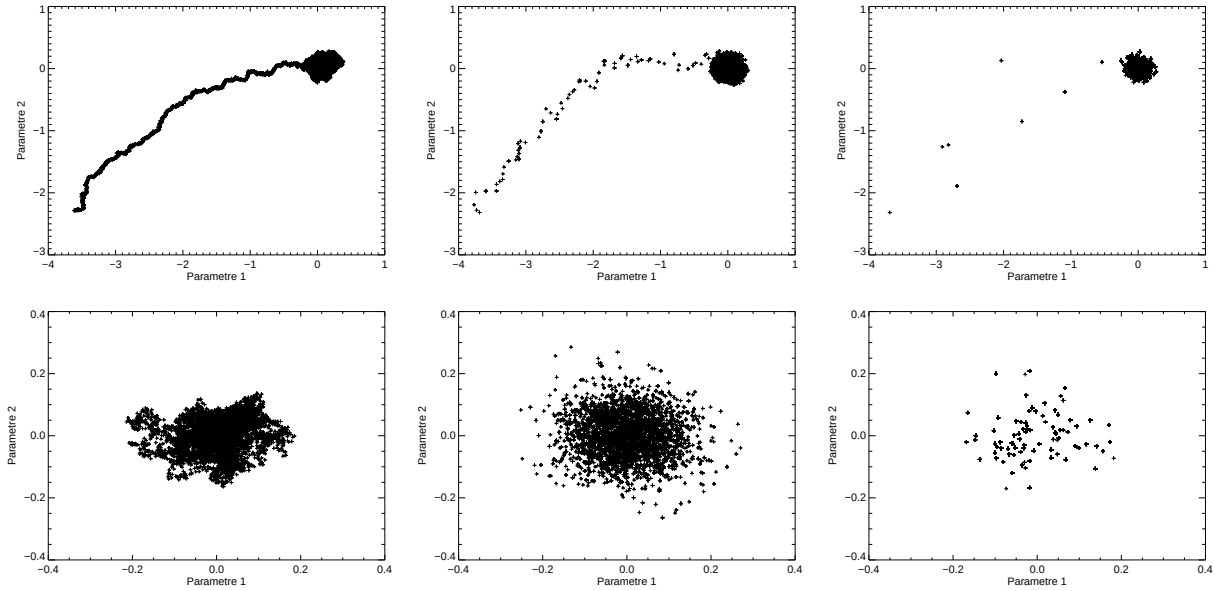


FIGURE 3.5 – Progression dans l'espace des paramètres (2D) d'une chaine de Markov pour 3 largeurs de la loi de proposition. A gauche, (rapport $1/10^{ème}$), la chaine progresse très lentement dans l'espace des paramètres car l'effet mémoire est important (forte corrélation entre points successifs de la chaine) et l'échantillonnage n'est pas homogène. Au centre (rapport 1), l'échantillonnage de la gaussienne est bien meilleur. A droite (rapport 10) la chaine reste longtemps à la même position dans l'espace des paramètres (effet plus visible sur la Fig.3.6) : le nombre d'échantillons indépendants est donc relativement faible.

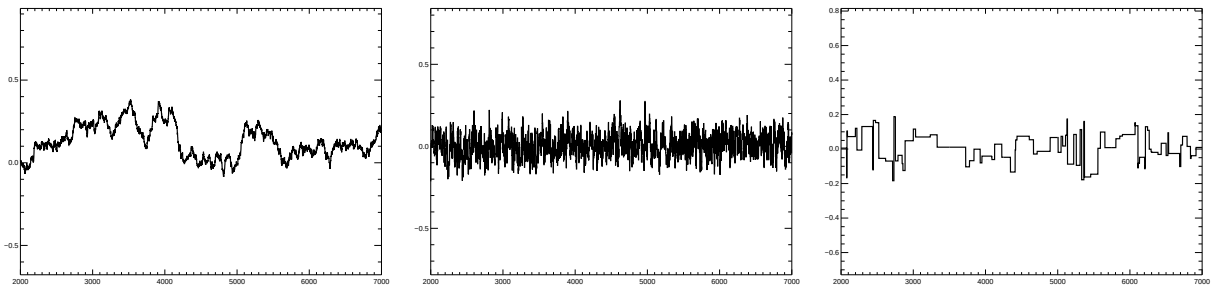


FIGURE 3.6 – Progression dans l'espace des parametres (2D) d'une chaine de Markov pour 3 largeurs de la loi de proposition en fonction des itérations (ici pris sur une section longue de 5000 échantillons). A gauche (rapport $1/10^{ème}$), les échantillons sont fortement corrélés entre eux. Au milieu (rapport 1), l'échantillonnage est homogène. A droite (rapport 10), la chaine reste longtemps à une position donnée et le nombre d'échantillons indépendants est faible (forte corrélation).

l'espace des paramètres. Cette vitesse est elle même reliée au *taux d'acceptation* (ou *acceptance rate*) de la chaine de Markov. Celui-ci est défini comme le nombre de transitions acceptées sur le nombre de transitions totales. Dans ces conditions, Roberts et al. (1997) ont montré que l'efficacité est maximale lorsque le *taux d'acceptation* est voisin de 23.4% pour $d \to \infty$ et de 44% pour des problèmes avec $d = 1$. Ce taux est atteint lorsque la matrice de covariance de la loi de proposition est de la forme $\sigma_q^2 I_d = (\phi^2/d) I_d$, I_d étant une matrice identité.

L'efficacité d'une chaine de Markov peut être estimée en déterminant le temps caractéristique pour que la chaine "oublie" ses états précédents. Idéalement, l'effet mémoire devrait être limité à 1 échantillon. Dans la pratique, la chaine de Markov reste un certain temps confinée dans un petit volume d'espace et la position à une itération i dépend de la position $i-1$ mais aussi légèrement des positions $i-2, i-3, i-k, ..., i-N$ (ce qui ne viole pas la propriété d'ergodicité puisque celle-ci est établit lorsque $N \to \infty$). Comme déjà suggéré précédemment, une manière d'estimer l'effet mémoire de la chaine consiste à calculer la corrélation entre un instant i et les instants précédents (après suppression de la phase de chauffage car naturellement, les positions sont fortement corrélées/directionnelles lorsque l'on progresse vers le maximum, cf. Fig.3.5),

$$\rho(k) = \frac{1}{N} \sum_{k=0}^{N} (x_i - \mu)(x_{i-k} - \mu) \tag{3.12}$$

Théoriquement, nous nous attendons à une fonction de corrélation exponentielle,

$$\rho(k) = e^{-\left|\frac{k}{\lambda}\right|} \tag{3.13}$$

où λ est un nombre caractéristique d'échantillons dont se "souvient" la $i^{ème}$ itération. L'efficacité de la chaine, définie comme la fraction d'échantillon indépendant, est estimé par l'expression,

$$E = \frac{1}{1 + 2\lambda}. \tag{3.14}$$

Si tous les échantillons sont indépendants, il n'y pas d'effet mémoire et l'efficacité est de 100%. Si le processus d'exploration ne dépend en moyenne que de l'iteration précédente, alors l'efficacité est de 33%. C'est donc l'efficacité maximale que peut avoir une chaine de Markov.

La figure 3.7 représente les fonctions de corrélation pour les trois scénarii précédemment présentés. Les valeurs du nombre d'itérations caractéristique de relaxation λ sont de 380, 6.2 et 25.6 pour respectivement des rapports d'écart types de 0.1, 1 et 10. Ceci correspond respectivement à des efficacités de 0.26%, 13.9% et 3.8%. Il semble donc plus préjudiciable d'explorer l'espace des parametres en faisant de petits "pas" plutôt que de grands "pas" dans l'espace des paramètres à faible dimension.

Pour des dimensions de l'espace des paramètres croissantes, l'efficacité de l'échantillonnage décroit : le processus de diffusion s'effectue dans un volume d'espace toujours plus grand. Ainsi, pour conserver une densité volumique d'échantillon constante, il faut augmenter leur nombre. Pour fixer les idées, la fraction d'échantillons indépendant passe de 20% à 1 dimension à 1% pour 129 paramètres [4] (cf. Fig.3.8). On peut reprendre l'exemple que nous avons utilisé précédemment pour le montrer. Nous avons modifié le modèle précité par un peigne de gaussiennes. Le procédé pour créer les données synthétiques est le même que celui décrit plus haut et la hauteur est encore une fois fixée.

En jouant sur le rapport entre les termes de variances de la matrice de covariance de la loi de proposition et les variances des variables, nous avons construit la figure 3.8, représentant

4. Cela signifie qu'il faut au moins 20 fois plus d'échantillons avec 129 paramètres qu'à une dimension pour obtenir un échantillonage de qualité égale.

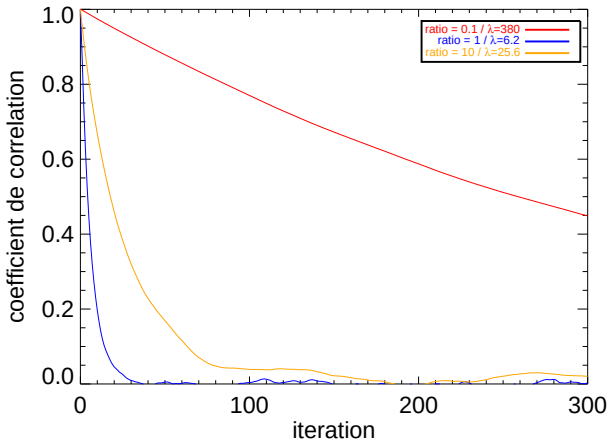


FIGURE 3.7 – Degré de corrélation entre la $N^{ème}$ itération d'une chaine de Markov et toutes les itérations précédentes, calculés pour 3 rapports de largeur entre la loi de proposition et de la distribution visée (appelé ratio ici).

l'efficacité en fonction du taux de réjection et de la taille de l'espace. Le comportement observé de l'efficacité est bien celui décrit par divers auteurs de la littérature (Roberts & Rosenthal 2001).

En règle générale, lorsque la dimensionnalité du problème est grande, il est admis que l'efficacité d'échantillonnage est optimale pour un taux d'acceptation compris entre 15% à 50%, car l'efficacité est relativement constante sur cette plage. Cependant, ces estimations ne sont dans l'absolu valides que dans le cas de lois de proposition et d'intérêt de même forme avec des variables non corrélées (comme c'est le cas dans nos exemples). Dans Roberts & Rosenthal (2001), une étude bien plus approfondie que la notre a été faite. Les auteurs précisent que le taux d'acceptation idéal peut varier légerement si, par exemple, la fonction d'intérêt comporte des variables discretes ou si celle ci est constituée de multiples maxima. Il faut aussi souligner que plus la loi de proposition est différente de la distribution d'intérêt (par exemple si on echantillonne une gaussienne avec une loi de Poisson), plus l'efficacité baisse. Cependant, dans la très grande majorité des cas explorés par divers auteurs, la plage optimale du taux de rejection est très peu affectée. En conclusion, on va supposer durant notre travail que la plage du taux d'acceptation jugée optimale, l'est aussi dans les problèmes qui nous concernent.

Néanmoins, nous sommes à peine plus avancés que précédemment car trouver le taux optimal de réjection nécessite que nous ayons une idée précise de la variance des paramètres. Ce qui, à l'exclusion des cas d'écoles, arrive très rarement. De multiples stratégies ont été avancées depuis l'explosion des techniques MCMC. La plupart du temps elles nécessitent une adaptation manuelle de la loi de proposition sur un principe essai/erreur. Mais dans les problèmes multi-variés de grande dimension, l'ajustement manuel relève du tour de force. Ces dernières années ont vu fleurir une nouvelle classe d'algorithmes MCMC qui constituent à nos yeux la meilleure solution. Cette classe d'algorithmes est appelé "Adaptive Markov Chain Monte-Carlo" (AMCMC). Les AMCMC ont la particularité d'optimiser l'efficacité d'une chaine de Markov en adaptant le taux d'acceptation en continu durant l'échantillonnage.

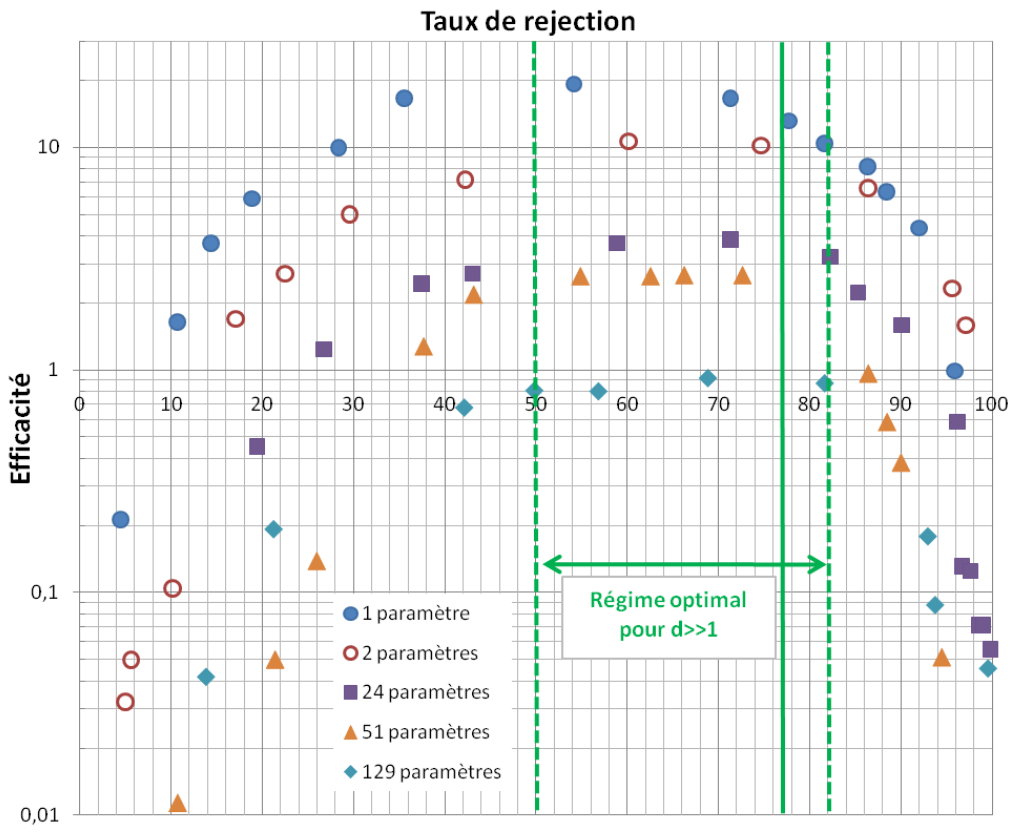


FIGURE 3.8 – Evolution de l'efficacité de l'algorithme en fonction du taux de réjection pour différents nombre de paramètres sur la base des mesures de λ que nous avons effectuées dans le cas d'un peigne de fonction gaussienne. Les taux optimum signalés dans la littératures sont retrouvés. On remarque aussi une tendance à la baisse de l'efficacité de l'algorithme de Metropolis en fonction du nombre de paramètres considérés.

3.2.3 L'Adaptive Markov Chain Monte-Carlo : un moyen de contrôle automatique du taux d'acceptation/réjection

Plusieurs auteurs se sont essayés à développer un AMCMC. L'un des plus récent a été développé par Atchadé (2006). Il utilise un algorithme stochastique de type Robbins-Monro lancé parallèlement à un algorithme de Metropolis-Hasting-Langevin (MALA pour *Metropolis Adjusted Langevin Algorithm*). Comparé à un algorithme de Metropolis-Hasting classique (appelé communément *Random Walk Metropolis* ou *RWM*), le MALA utilise le gradient local de la loi de proposition pour améliorer la vitesse de convergence vers le maximum. Pour permettre une adaptation facile de la loi de proposition, celle ci est prise gaussienne de matrice de covariance Σ. Cet algorithme, contrairement à la plupart de ses prédécesseurs, ne présuppose pas la matrice comme étant diagonale. En conséquence, après la *phase d'apprentissage* de la matrice de covariance, les éventuels termes de corrélation sont pris en compte, ce qui peut accélérer l'échantillonnage. Nous nous sommes fortement inspirés de l'algorithme présenté par Atchadé pour écrire notre code en langage IDL et allons donc le décrire dans les grandes lignes.

A notre connaissance, les performances de tels algorithmes n'ont pas été intensivement testées dans les problèmes de grandes dimensions. Or ce sont ceux ci qui nous intéressent au plus haut point dans les problèmes d'analyse de données astérosismiques. De plus, il est signalé par Atchadé que le MALA a une tendance à légèrement sous-estimer la variance des fonctions de densité de probabilité car le terme de gradient a tendance à confiner l'échantillonnage autour du maximum local de la distribution d'intérêt. Enfin, nous n'avons trouvé dans la littérature aucune analyse du comportement du MALA dans des situations où la loi d'intérêt est multimodale. Nous soupçonnons que le terme de gradient peut nuire à l'échantillonnage dans ce cas, en forçant la chaine de Markov à préférentiellement converger vers le maximum le plus proche du point de départ.

Notre objectif n'étant pas d'effectuer une analyse des propriétés statistiques complète des algorithmes MCMC, mais des les utiliser en tant qu'outils d'analyse des modes de pulsations stellaires, nous avons préféré la prudence en restant en terrain connu.

En conséquence, bien que nous ayons écrit notre code sur le modèle algorithmique proposé par Yves Atchadé, nous l'avons simplifié pour l'utiliser avec un RWM, bien mieux connu que le MALA. Dans ce qui suit, nous montrerons où cette simplification s'effectue.

Définitions

L'approche AMCMC de Y. Atchadé consiste à ajouter aux paramètres d'intérêts $\boldsymbol{\theta}$ un ensemble d'hyperparamètres $\boldsymbol{\Omega}$ chargé de l'ajustement de la matrice de covariance. Ce dernier est défini par 3 objets mathématiques :

- Un vecteur $\boldsymbol{\mu}$ de dimension N identique à celle de $\boldsymbol{\theta}$, tel que $|\boldsymbol{\mu}| \leq A_1$. $\boldsymbol{\mu}$ est donc contenue dans une hypersphère de rayon A_1.
- Une matrice définie semi-positive $\boldsymbol{\Gamma}$ (*i.e.*, $\langle \boldsymbol{x}| \boldsymbol{\Gamma} |\boldsymbol{x}\rangle \geq 0$) dont la norme (au sens de Frobenius) est borné : $|\boldsymbol{\Gamma}| \leq A_1$. Sa dimension est $N \times N$.
- Un scalaire σ tel que $\sigma \in [\epsilon_1, A_1]$.

On a donc $\boldsymbol{\Omega} = \{\boldsymbol{\mu}, \boldsymbol{\Gamma}, \sigma\}$. ϵ_1 et A_1 sont des bornes dont le rôle sera explicité plus loin. Pour rappel, la norme de Frobenius est issue du produit scalaire $A \cdot B = \sqrt{\sum_{i,j} A_{i,j} B_{i,j}}$. $\boldsymbol{\mu}$ correspond à la moyenne empirique de $\boldsymbol{\theta}$. Plus précisément, asymptotiquement, lorsque l'itération $t \to \infty$, $\mu_{(t)}$ doit être égale à la moyenne des échantillons de $\boldsymbol{\theta}$. Pour éviter la confusion entre indice des éléments des vecteurs et index des échantillons, ces derniers seront systématiquement notés entre parenthèses. De la même manière, $\boldsymbol{\Gamma}_{(t)}$ est reliée à la covariance empirique des échantillons de

chacune des variables et $\sigma_{(t)}$ est un facteur d'échelle modifiant la matrice de covariance de la loi de proposition. Celle ci est constituée d'une matrice de covariance construite à chaque itération selon le schéma,

$$\boldsymbol{\Sigma_{(t)}} = \sigma^2_{(t)}(\Gamma_{(t)} + \epsilon_2 \boldsymbol{I_d}) = \sigma^2_{(t)}\boldsymbol{\Lambda_{(t)}}, \tag{3.15}$$

où ϵ_2 est un scalaire définissant la quantité caractéristique à utiliser pour réviser la matrice de covariance par une petite matrice diagonale $\epsilon_2 \boldsymbol{I_d}$. Typiquement $\epsilon_2 = 10^{-8}$. Cette étape ne correspond pas à l'apprentissage. La matrice diagonale ajoutée ne sert qu'à s'assurer que la matrice est toujours inversible[5] (donc défini semi-positive). En effet, les arrondis numériques peuvent aboutir à des valeurs négatives sur la diagonale, ce qui est à proscrire. On s'attend à ce que pour $t \to \infty$, $\boldsymbol{\Lambda_{(t)}}$ soit égale à la matrice des variances-covariances de la loi d'intérêt $\pi(\boldsymbol{\theta})$. On a omis les *a priori* et la notation conditionnelle pour alléger l'écriture.

Nommons $D(\boldsymbol{\theta_{(t)}})$ la fonction contenant le terme de gradient local propre au MALA à la $t^{ème}$ itération du processus d'échantillonnage. On l'appelle fonction de dérive (*drift function*) car elle a pour rôle d'orienter l'échantillonage afin d'accélérer la convergence vers la zone du maximum de probabilité. Dans le cas d'un MALA (plus précisément un MALA tronqué, c'est à dire pour lequel on limite l'influence du terme de gradient) une expression couramment utilisé de la fonction de dérive est

$$D(\boldsymbol{\theta_{(t)}}) = \frac{\delta}{max\left\{\delta, \left|\nabla ln\, \pi(\boldsymbol{\theta_{(t)}})\right|\right\}} \nabla ln\, \pi(\boldsymbol{\theta_{(t)}}). \tag{3.16}$$

δ définit le seuil à partir duquel on commence à limiter le terme de gradient $\nabla ln\, \pi(\boldsymbol{\theta_{(t)}})$. La loi de proposition s'écrit formellement,

$$q(\boldsymbol{\theta_{(t)}}) = N\left(\boldsymbol{\theta_{(t)}} + \frac{\boldsymbol{\Sigma_{(t)}}}{2} D(\boldsymbol{\theta_{(t)}}),\, \boldsymbol{\Sigma_{(t)}}\right) \tag{3.17}$$

Ce qui signifie que la mise à jour des paramètres s'effectue selon le schéma,

$$\boldsymbol{\theta_{(t+1)}} = \boldsymbol{\theta_{(t)}} + \frac{\boldsymbol{\Sigma_{(t)}}}{2} D(\boldsymbol{\theta_{(t)}}) + N(0,\, \boldsymbol{\Sigma_{(t)}}). \tag{3.18}$$

On fera remarquer que lorsque $D(\boldsymbol{\theta_{(t)}}) \equiv 0$ (imposé nul), l'algorithme MALA est strictement équivalent au RWM Metropolis-Hasting.

De manière similaire au Chap.3.2.1, $\alpha(\boldsymbol{\theta_{(t+1)}}, \boldsymbol{\theta_{(t)}})$ désignera la probabilité de transition[6] entre la position t et $t+1$. Lorsque $D(\boldsymbol{\theta_{(t)}}) \equiv 0$, la loi de proposition (gaussienne) est symétrique autrement dit, $q(\boldsymbol{\theta_{(t+1)}}|\boldsymbol{\theta_{(t)}}) = q(\boldsymbol{\theta_{(t)}}|\boldsymbol{\theta_{(t+1)}})$. L'algorithme est dit de Metropolis et l'Eq.3.9 s'écrit dans notre contexte,

$$\alpha(\boldsymbol{\theta_{(t+1)}}, \boldsymbol{\theta_{(t)}}) = min\left\{1, \frac{\pi(\boldsymbol{\theta_{(t+1)}})}{\pi(\boldsymbol{\theta_{(t)}})}\right\} \tag{3.19}$$

Cette situation particulière correspond à l'algorithme de Metropolis originel, avant que Hasting ne le généralise à n'importe quelle classe de loi de proposition.

Notre implémentation se place dans cette condition particulière où la fonction de dérive $D(\boldsymbol{\theta_{(t)}})$ n'est pas pris en compte. Dans la suite nous ne reparlerons plus de ce terme : sa valeur sera implicitement nulle pour alléger l'écriture. On soulignera que si à l'avenir, nous souhaitions changer d'avis, les modifications à effectuer sur le code sont mineures.

5. La matrice des covariances doit être inversé pour construire la comatrice des covariances et permettre dans la pratique le tirage aléatoire.

6. Il ne faut pas oublier que cet probabilité est ici dépendante des hyperparamètres $\boldsymbol{\sigma_{(t)}}$ et $\boldsymbol{\Lambda_{(t)}}$.

Présentation de l'Approximation Stochastique de Robbins-Monro

Avant d'expliciter l'algorithme adaptatif que nous avons utilisé, il convient de décrire ce qu'est un algorithme de *gradient stochastique* de Robbins-Monro (Robbins & Monro 1951). Robbins H. et Monro S. ont proposé un schéma itératif pour rechercher les solutions d'une fonction bruitée. L'idée consiste à rechercher le ou les zéro(s) d'une fonction $H(\beta, x)$ qui dépend du paramètre β et bruitée par la variable aléatoire x. Ici on recherche plus exactement $< H(\beta_0, x) >_x = 0$, c'est à dire la position moyenne de l'intersection avec l'abscisse de la fonction. Cette recherche s'effectue grâce au schéma numérique itératif,

$$\beta_{(t+1)} = \beta_{(t)} + \gamma_{(t)} H(\beta_{(t)}, x_{(t)}) \tag{3.20}$$

qui est un schéma numérique possible de l'équation,

$$\frac{d\beta(t)}{dt} = H(\beta(t), x(t)). \tag{3.21}$$

Proche de la solution $\beta \approx \beta_0$ et,

$$\frac{d\beta(t)}{dt} \simeq 0. \tag{3.22}$$

Un des gros avantages de ce type d'algorithme est qu'il est séquentiel et donc utilisable sur des données acquises en temps réel : seule la dernière valeur peut-être retenue et toutes les autres oubliées.

Le caractère stochastique de la recherche est dû à la nature bruitée des valeurs x_t, ce qui justifie le nom de recherche par gradient stochastique. $\gamma_{(t)}$ est un scalaire dont la valeur décroit ni trop lentement ni trop rapidement (à ajuster manuellement) de manière monotone d'une itération t sur l'autre. C'est le *coefficient d'amortissement du processus d'apprentissage*. Les propriétés de convergence sont étroitement liées à ce dernier paramètre. Il faut que,

$$lim_{N\to\infty}\gamma_{(t)} = 0 \tag{3.23}$$

$$\sum_{N=1}^{\infty} \gamma_{(t)} = \infty \tag{3.24}$$

$$\sum_{N=1}^{\infty} \gamma_{(t)}^2 < \infty. \tag{3.25}$$

La première condition (Eq.3.23) garantit que les corrections apportées par le terme $H(\beta_{(t)}, x_{(t)})$ décroissent en magnitude de telle sorte que le processus converge vers une valeur finie. La seconde condition (Eq.3.24) garantit que le coefficient d'amortissement est suffisamment grand pour qu'une solution soit trouvée. Enfin la dernière condition (Eq.3.25) permet de s'assurer que les fluctuations stochastiques de $H(\beta_{(t)}, x_{(t)})$ sont modérées (bruit de variance fini) de telle sorte que celles ci ne viennent pas perturber trop fortement la solution stationnaire à l'arrêt de l'apprentissage.

Au fur est à mesure que le terme $\gamma_{(t)}$ décroit, le terme de perturbation $\gamma_{(t)} H(\beta_{(t)}, x_{(t)})$ (de moyenne selon $\beta_{(t)}$ et $x_{(t)}$ nulle) est réduit et l'algorithme converge vers la solution recherchée. Asymptotiquement, lorsque $\gamma_{(t)} \to 0$, nous atteignons l'état stationnaire pour β, solution de l'optimisation. On a représenté schématiquement la manière dont converge le processus sur la Fig.3.9.

Les équations 3.31a, 3.31b, 3.31c de la phase d'apprentissage de notre algorithme utilisent l'algorithme de Robbins-Monro. Par exemple, pour ajuster le paramètre σ à sa valeur optimale σ_{opt}

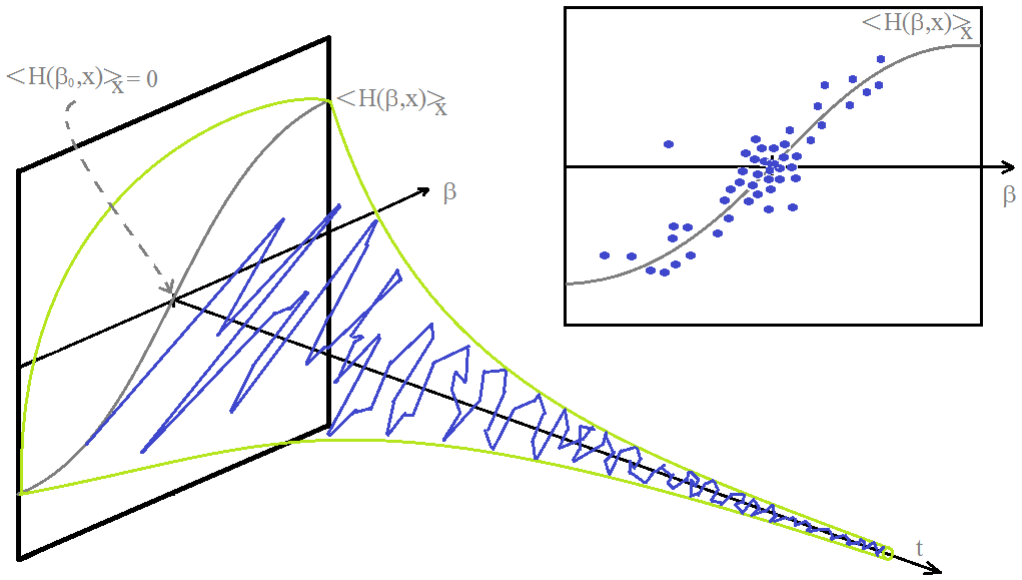


FIGURE 3.9 – Représentation schématique de la façon dont converge un algorithme de Robbins-Monro vers le zéro du critère $H(\beta, x)$. Sur la gauche on a représenté la dépendance de $H(\beta, x)$ vis à vis de t et de β. L'enveloppe verte délimite le volume d'espace exploré au fur et à mesure que croit le nombre d'échantillons t. La surface limite du volume est contrôlée par γ_t. Asymptotiquement la valeur converge vers l'optimum défini par les coordonnées $\{\beta_0, < H(\beta_0, x) >_x = 0\}$. L'encadré représente une projection de la figure de gauche. Les différentes valeurs obtenues des échantillons sont dispersées autour de la fonction $< H(\beta, x) >_x$ du fait du caractère stochastique de x. Si les conditions de convergence sont remplies, lorsque $t >> 0$, les échantillons seront concentrés autour de l'optimum. Dans ce type de processus, seule la dernière valeur obtenue est retenue.

on utilise la fonction coût $H(\boldsymbol{\sigma_{(i+1)}}, \boldsymbol{\theta_{(i+1)}}) = \alpha(\boldsymbol{\theta_{(i+1)}}, \boldsymbol{\theta_{(i)}}) - \tau$, τ étant le taux idéal d'acceptation des échantillons[7]. La dépendance à σ du terme de droite est cachée dans $\alpha(\boldsymbol{\theta_{(i+1)}}, \boldsymbol{\theta_{(i)}})$. Ici c'est l'algorithme MCMC qui fournit les valeurs aléatoires $x \equiv \boldsymbol{\theta}$.

On voit ici que l'on cherche à réduire l'écart entre la valeur recherchée τ et la valeur mesurée $\alpha(\boldsymbol{\theta_{(i+1)}}, \boldsymbol{\theta_{(i)}})$. Ceci sera rappelé lorsque sera présenté l'algorithme d'apprentissage.

Ce type d'algorithme stochastique a été utilisé par plusieurs auteurs pour optimiser automatiquement les AMCMC. Il convient de préciser que le couplage d'un algorithme de Robbins-Monro avec un MCMC ne garantit pas forcément la propriété d'ergodicité du MCMC. Autrement dit, l'ajout de l'hyperparamètre β modifie les propriétés de convergence de la chaine de Markov : la solution stationnaire n'est plus forcément la distribution d'intérêt. La grande majorité des auteurs recommandent donc l'arrêt de la phase d'apprentissage au bout d'un nombre d'itération suffisamment élevé (à déterminer empiriquement). Une description plus complète de la méthode d'approximation stochastique appliquée à diverses MCMC peut être trouvée dans Andrieu & Robert (2001).

L'algorithme

Le module chargé de la réactualisation des valeurs des hyper-paramètres $\boldsymbol{\Omega}$ répond à l'algorithme suivant.

A une itération j du processus d'échantillonnage, on initialise le jeu d'hyper-paramètres, $\boldsymbol{\Omega_{(j)}} = \left\{\boldsymbol{\mu_{(j)}}, \boldsymbol{\Gamma_{(j)}}, \sigma_{(j)}\right\}$. Puis pour toutes les itérations i suivantes ($i \geq j$) on procède comme suit,

1/ On incrémente la matrice $\boldsymbol{\Lambda_{(i)}}$ pour améliorer la stabilité de l'algorithme (matrice toujours inversable),

$$\boldsymbol{\Lambda_{(i)}} = \boldsymbol{\Gamma_{(i)}} + \epsilon_2 \boldsymbol{I_d}. \tag{3.26}$$

2/ On propose une nouvelle position $\boldsymbol{Y_{(i+1)}}$ dans l'espace des paramètres, générée en fonction de l'ancienne $\boldsymbol{\theta_{(i)}}$ et de la *loi de proposition* $q(\boldsymbol{\theta_{(i)}})$,

$$\boldsymbol{Y_{(i+1)}} = \boldsymbol{\theta_{(i)}} + N(0, \sigma^2_{(i)} \boldsymbol{\Lambda_{(i)}}). \tag{3.27}$$

3/ On exécute l'algorithme de Metropolis-Hasting proprement dit.

► On génère un nombre aléatoire u dont la distribution statistique est uniforme,

$$u \sim U[0,1] \tag{3.28}$$

► Si $u \leq \alpha_{\sigma_{(i)}, \Lambda_{(i)}}(\boldsymbol{Y_{(i+1)}}, \boldsymbol{\theta_{(i)}})$ alors on accepte la position proposée,

$$\boldsymbol{\theta_{(i+1)}} = \boldsymbol{Y_{(i+1)}} \tag{3.29}$$

Sinon on la rejette,

$$\boldsymbol{\theta_{(i+1)}} = \boldsymbol{\theta_{(i)}} \tag{3.30}$$

4/ On réactualise chacun des hyperparamètres contrôlant l'efficacité de l'échantillonnage selon un schéma de Robbins-Monro[8]. C'est la séquence d'apprentissage proprement dite,

$$\boldsymbol{\mu_{(i+1)}} = p_3(\boldsymbol{\mu})(\mu_{(i)} + \gamma_{(i)}(\boldsymbol{\theta_{(i+1)}} - \boldsymbol{\mu_{(i)}})) \tag{3.31a}$$

7. On rappelle que le terme σ est le paramètre d'échelle de la matrice de covariance $\boldsymbol{\Sigma_{(t)}} = \sigma^2_{(t)} \boldsymbol{\Lambda_{(t)}}$.

8. Avec la notation de Dirac, construire un espace vectoriel se note $|\boldsymbol{x}\rangle \langle \boldsymbol{y}|$. Par exemple avec $\boldsymbol{x} = \boldsymbol{y} = (a, b)$ on obtient une matrice carré : $|\boldsymbol{x}\rangle \langle \boldsymbol{y}| = \begin{pmatrix} a^2 & ab \\ ab & b^2 \end{pmatrix}$

$$\boldsymbol{\Gamma_{(i+1)}} = p_2(\boldsymbol{\Gamma})(\Gamma_{(i)} + \gamma_{(i)}(\left|\boldsymbol{\theta_{(i+1)}} - \boldsymbol{\mu_{(i)}}\right\rangle \left\langle \boldsymbol{\theta_{(i+1)}} - \boldsymbol{\mu_{(i)}}\right| - \boldsymbol{\Gamma_{(i)}})) \tag{3.31b}$$

$$\sigma_{(i+1)} = p_1(\sigma)(\sigma_{(i)} + \gamma_{(i)}(\alpha(\boldsymbol{Y_{(i+1)}}, \boldsymbol{\theta_{(i)}}) - \bar{\tau})). \tag{3.31c}$$

5/ On incrémente i et on retourne à l'étape 1,

$$i \leftarrow i + 1 \tag{3.32}$$

L'étape 4 nécessite d'être décrite plus en profondeur. A celle-ci, l'algorithme de Robbins-Monro est utilisé pour minimiser l'écart entre les deux premiers moments (Eq. 3.31a et Eq. 3.31b) de la loi de proposition et de la distribution d'intérêt (moyenne et covariance). L'idée, relativement intuitive, consiste à essayer d'obtenir une approximation de la distribution d'intérêt. Comme signalé au Chap.3.2.2, la qualité de l'échantillonnage dépend aussi du taux d'acceptation. En conséquence, l'écart entre un taux "ciblé" supposé optimal $\bar{\tau}$ et la probabilité de transition, constitue le troisième critère utilisé par l'algorithme (Eq. 3.31c). Pour rappel, le taux optimal est $\bar{\tau} = 0.234$ dans le cas d'une distribution d'intérêt gaussienne. Plus généralement, on considère qu'un taux d'acceptation compris entre 0.5 et 0.18 est optimal.

On remarque aussi 3 termes, $p1(\sigma)$, $p_2(\boldsymbol{\Gamma})$, $p_3(\boldsymbol{\mu})$, qui n'ont pas encore été définis. Ce sont des fonctions permettant de borner l'espace d'exploration des hyperparamètres $\boldsymbol{\Omega}$. En se référant aux ensembles de définition décrits au Chap.3.2.3, on a,

$$\begin{aligned} p1(\sigma) &= \sigma \text{ si } \epsilon_1 \leq \sigma \leq A_1 \\ &= \epsilon_1 \text{ si } \sigma < \epsilon_1 \\ &= A_1 \text{ si } \sigma > A_1. \end{aligned} \tag{3.33}$$

$$\begin{aligned} p_2(\boldsymbol{\Gamma}) &= \boldsymbol{\Gamma} \text{ si } \boldsymbol{\Gamma} \leq A_1 \\ &= \frac{A_1}{|\boldsymbol{\Gamma}|}\boldsymbol{\Gamma} \text{ sinon.} \end{aligned} \tag{3.34}$$

$$\begin{aligned} p_3(\boldsymbol{\mu}) &= \boldsymbol{\mu} \text{ si } \boldsymbol{\mu} \leq A_1 \\ &= \frac{A_1}{|\boldsymbol{\mu}|}\boldsymbol{\mu} \text{ sinon.} \end{aligned} \tag{3.35}$$

Les bornes ϵ_1 et A_1 sont en fait peu contraignantes mais nécessaires pour garantir l'ergodicité de l'algorithme (ou tout du moins sa démonstration). On s'assure donc que la borne A_1 est suffisamment grande (typiquement on a pris $A_1 = 10^{10}$) et ϵ_1 suffisamment petit ($\epsilon_1 = 10^{-8}$).

Lors des différentes utilisations que nous avons fait de l'algorithme, nous avons défini $\bar{\tau} = 0.27$. L'approximation obtenue du taux d'acceptation n'est pas toujours parfaite : elle dépend de la valeur initiale de la constante de couplage $c_0 = \gamma_{i=0}$, de la vitesse de décroissance choisie de γ_i et du nombre d'itérations ayant servi à l'apprentissage (idéalement infini). Y. Atchadé suppose que d'une manière générale, on peut prendre $\gamma_i = c_0/i^\lambda$, avec $\lambda \in [1/2, 1]$ et $c_0 > 0$ de telle sorte que $\sum_i \gamma_i = \infty$ (cf. condition de convergence de l'algorithme de Robbins-Monro, Eq.3.24). Il recommande cependant d'utiliser $\lambda = 1$, recommandation que nous avons suivie.

En fonction du nombre de paramètres d'intérêt que nous utilisions, nous avons remarqué qu'il était nécessaire d'ajuster c_0. Typiquement, nous utilisons $c_0 = 2$ pour des problèmes de grande dimension, (*i.e.* $dim(\boldsymbol{\theta}) \gtrsim 30$) et plutôt $c_0 = 5$ à $c_0 = 10$ dans les cas de moindre dimension.

On fera remarquer que l'ergodicité est vérifiée pour cet algorithme. Néanmoins, nous avons tout de même préféré découpler la phase d'apprentissage de la phase d'échantillonnage et ce pour une raison assez simple. Nous avons constaté qu'à nombre d'échantillons égal, l'ajout des hyper-paramètres de contrôle du taux d'échantillonnage confinait la chaine de Markov au voisinage du maximum de la distribution d'intérêt (sous estimation de la variance des paramètres). Ce qui

signifie que l'exploration de l'espace des paramètres se fait bien plus lentement (Cf. Chap.3.2.2, partie traitant de l'efficacité des MCMC). En effet, nous ajoutons un nombre important de paramètres à ajuster. Typiquement, pour un problème de dimension 50, le nombre total d'hyper-paramètres indépendant est de l'ordre de la centaine[9]. Une fois le taux optimal raisonnablement approché, il convient donc de couper l'apprentissage.

Il est inutile durant les premières itérations de lancer l'apprentissage. En effet, on se situe en général initialement assez loin du maximum de probabilité et le taux d'acceptation est anormalement élevé au début. Typiquement, nous ne commençons la phase d'apprentissage qu'au bout de $j = 800$ itérations. Le temps auquel l'apprentissage est arrêté dépend du nombre de paramètres (valeur typique de 1 à 1.5 millions d'itérations pour 80-120 variables).

Pour conclure, l'algorithme adaptatif d'échantillonnage de Monte-Carlo par chaine de Markov présenté ici permet de réduire drastiquement le nombre de paramètres de contrôle de la chaine de Markov. En effet, nous avons substitué la matrice de la *loi de proposition* contenant $N \times N$ éléments par 1 unique paramètre de contrôle : la constante de couplage c_0. Même si le temps d'apprentissage nécessaire peut se révéler très long, dans les problèmes de très grande dimension N, le gain de temps comparé à un ajustement manuel de la matrice de covariance est considérable.

Il faut toutefois être relativement prudent sur le résultat de l'apprentissage car si la *loi d'intérêt* est fortement multimodale avec des maxima éloignés, alors l'apprentissage peut être imparfait. En effet, prenons l'exemple d'un critère statistique constitué de 2 maxima locaux fortement prononcés et disjoint. Une fois que la chaine de Markov a convergé vers l'un ou l'autre des maxima de probabilité, l'apprentissage va se faire en accord avec la variance dudit maximum local. Malheureusement dans ce cas, il est très fortement probable que la probabilité de transition d'un maximum à l'autre soit quasi nulle. Il faut donc trouver un moyen de faciliter le transit entre maxima.

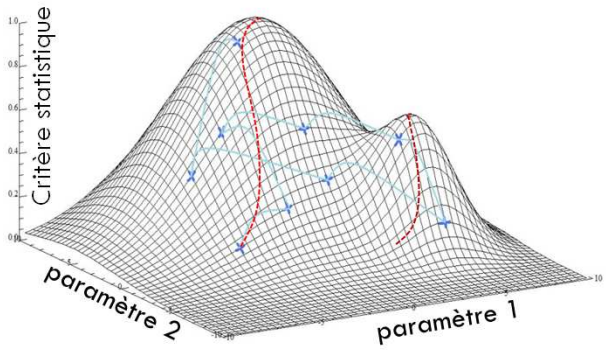


FIGURE 3.10 – Exemple de critère statistique avec 2 maxima de densité de probabilité, dans un espace 2D. Le cas d'un échantillonnage par chaine de Markov est représenté par des croix rouges et arcs bleus. Un processus de recherche du maximum par descente de gradient est représenté par la ligne pointillé rouge. Ce dernier convergera vers l'un ou l'autre des maxima en fonction du vecteur initial de paramètre. L'efficacité d'un algorithme de Metropolis-Hasting peut être mis à rude épreuve si les deux maxima de densité de probabilité sont très éloignés.

9. Le nombre d'hyperparamètres indépendant est donné par $Dim(Trace(\Gamma)) + Dim(\mu) + 1$. Les termes non-diagonaux de la matrice ne sont pas indépendants puisqu'ils sont obtenus par des produits croisés.

3.2.4 Eviter le confinement dans des maxima locaux du critère statistique

Plusieurs solutions ont été proposées pour éviter le confinement de la chaine dans un maximum local mais c'est sans conteste le *parrallel tempering* (recuit parallèle en français) qui a eu le plus de succès. Cette approche a été introduite pour la première fois par Swendsen & Wang (1986) et formalisé par Geyer (1991) et Jennison (1993), elle emprunte sa terminologie à la physique car l'idée originale était fortement inspirée de la physique/chimie atomique (modèle de spin d'Ising par exemple) et de nos jours couramment appliquée en biologie moléculaire (modèles de protéines). Pour comprendre la philosophie du *parrallel tempering* et familiariser le lecteur à cette notion, nous allons nous servir de deux exemples analogues issus de la physique/chimie/biologie et où les TMCMC (pour *Tempered MCMC*) sont couramment utilisés.

Etats ro-vibrationnels d'une molécule. Une molécule peut être vue comme un ensemble organisé d'atomes en interaction. Ce système physique peut vibrer, subir des rotations/torsions ou se disloquer. Les molécules disposent donc de plusieurs états stables possibles correspondant à autant de configurations spatiales. Les transitions entre états se font par absorption/ré-émission de quantum d'énergie. Ces différents états ro-vibrationnels sont quantifiés et les différentes configurations spatiales d'une molécule peuvent parfois être séparées par de forte barrières de potentiel. C'est notamment le cas des protéines où à certaines températures critiques d'un gaz moléculaire, des transitions de phases se produisent, correspondant à des changements de géométrie moléculaires majeurs. La dislocation d'une molécule n'est qu'une de ces transitions de phase possibles. Les biologistes utilisent couramment les chaines de Markov tempérées pour mettre en évidence les différents états stables possibles des protéines en fonction de la température (voir par exemple Lin et al. (2003)), car l'application des chaines de Markov est simple et intuitive dans ce domaine.

Modèle d'Ising. Le modèle d'Ising est un modèle simple permettant d'expliquer (entre autre) les propriétés magnétiques de certains matériaux. On considère un réseau de dipôles magnétiques ne pouvant se trouver que dans deux configurations possibles, caractérisés par des moments magnétiques $+M$ et $-M$. En dessous d'une température critique T_c ce réseau verra tous ses dipôles orientés dans la même direction ($+M$ ou $-M$) et le matériaux est alors ferromagnétique (présence d'un champ magnétique permanent en l'absence de toute influence extérieure). Au delà, à cause de l'agitation thermique, les dipôles s'orientent aléatoirement et le moment magnétique global du matériau est nul en l'absence d'influence extérieure. Comme expliqué plus loin, ce type de transition peut être directement mis en évidence par le *parrallel tempering*.

Avec ces descriptions, on se rend compte que les transitions entre état éloignés du système d'intérêt sont plus faciles à haute température qu'à basse température : La température abaisse les barrières de potentiel (Cf. Fig.3.11). Dans le cas du modèle d'Ising, elle permet un mélange des états individuels des dipôles augmentant de manière abrupte (effet de seuil) l'entropie. Or, lors d'un échantillonnage d'une *loi d'intérêt* nous cherchons à échantillonner *toute* la fonction et donc, nous souhaitons nous affranchir des barrières de potentiels éventuelles. L'idée maîtresse consiste donc à *altérer la distribution d'équilibre* en introduisant une dépendance à la température. Le *parrallel tempering* consiste alors à intervertir continuellement durant le processus d'échantillonnage, deux échantillons de deux chaines (généralement mitoyennes) k et l (choisies au hasard), avec une probabilité α_{switch}. Le fait d'intervertir de manière probabiliste deux chaines de températures différentes s'apparente à un effet tunnel en physique (Katzgraber et al. 2006). En reprenant la pensée de Metropolis (Cf. Chap.3.1.2), on écrira donc la probabilité d'intervertir

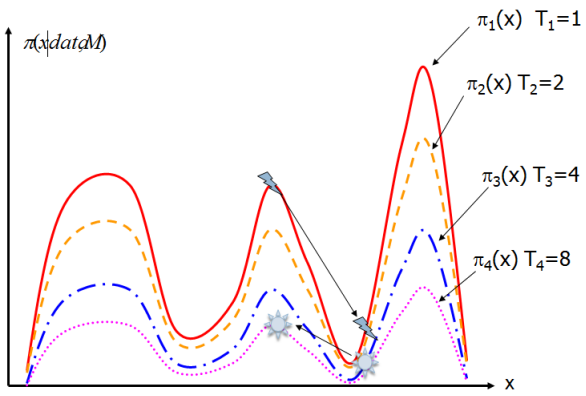


FIGURE 3.11 – Illustration de l'effet du *parrallel tempering* sur une distribution d'intérêt multimodale. Plus le paramètre température est élevé, plus on abaisse les barrières de potentiel entre les différents maxima de la distribution d'intérêt. Les transitions entre vallées et creux de la fonction coût sont plus faciles à haute température. le *parrallel tempering* consiste à se servir de cette propriété en mélangeant régulièrement les échantillons d'une chaine de température T_k avec ceux de la chaine voisine, plus chaude (T_{k+1}). Ce procédé de mélange s'apparente à un effet tunel.

deux positions $\{E_k, \beta_k\}$ et $\{E_l, \beta_l\}$ comme

$$p(\{E_k, \beta_k\} \longrightarrow \{E_l, \beta_l\}) = min\left[1, e^{\Delta\beta\Delta E}\right], \tag{3.36}$$

Avec β l'inverse de la température : $\beta_k = 1/T_k$. La température de la chaine de Markov devient donc une variable supplémentaire [10]. $\Delta\beta$ et ΔE sont définies comme,

$$\Delta\beta = \beta_l - \beta_k \tag{3.37}$$

$$\Delta E = E_l - E_k \tag{3.38}$$

Si on souhaite acquérir une vision plus formelle des analogies possibles entre physique et la technique du *parrallel tempering* on pourra se référer Earl & Deem (2005).

D'un point de vue statistique, on écrira que la probabilité d'accepter ou non un mélange de chaines à l'itération i entre deux chaines tempérées mitoyennes k et $k+1$ comme

$$\alpha_{switch}(\boldsymbol{\theta}^{\boldsymbol{k+1}}_{(\boldsymbol{i})}, \boldsymbol{\theta}^{\boldsymbol{k}}_{(\boldsymbol{i})}) = min\{1, \frac{\pi_k(\boldsymbol{\theta}^{\boldsymbol{k+1}}_{(\boldsymbol{i})}|y, M, I)\pi_{k+1}(\boldsymbol{\theta}^{\boldsymbol{k}}_{(\boldsymbol{i})}|y, M, I)}{\pi_k(\boldsymbol{\theta}^{\boldsymbol{k}}_{(\boldsymbol{i})}|y, M, I)\pi_{k+1}(\boldsymbol{\theta}^{\boldsymbol{k+1}}_{(\boldsymbol{i})}|y, M, I)}\}. \tag{3.39}$$

Moyennant les changement de variables adéquat l'Eq.3.39 est strictement équivalente l'Eq.3.36. Les termes de la forme $\pi_k(\boldsymbol{\theta}^{\boldsymbol{k}}_{(\boldsymbol{i})}|y, M, I)$ représentent les *lois d'intérêts tempérées* (*a posteriori*) utilisées lors de l'échantillonnage à température T_k. Dans la pratique, comme les *a priori* constituent l'information minimale dont on dispose, les distributions *a priori* associées ne doivent pas être affectées par la température. En somme, seul le terme de vraisemblance est

10. On rappelle que dans l'analogie utilisé par Metropolis, ΔE correspond à l'énergie nécessaire pour transiter d'un état à un autre.

tempéré de telle sorte que si on nomme la chaine la plus froide $T_1 = 1/\beta_1 = 1$, correspondant par définition à la distribution *a posteriori* d'intérêt ($\pi_{k=1}(\boldsymbol{\theta}^{\boldsymbol{k}}_{(\boldsymbol{i})}|y, M, I) = \pi(\boldsymbol{\theta}_{(\boldsymbol{i})}|y, M, I)$), on a

$$\pi_k(\boldsymbol{\theta}^{\boldsymbol{k}}_{(\boldsymbol{i})}|y, M, I) \propto \pi(y|\boldsymbol{\theta}^{\boldsymbol{k}}_{(\boldsymbol{i})}, M, I)^{\beta_k}\pi(\boldsymbol{\theta}_{(\boldsymbol{i})}|M, I). \tag{3.40}$$

A température infinie, la vraisemblance "chauffée" $\pi(y|\boldsymbol{\theta}^{\boldsymbol{k}}_{(\boldsymbol{i})}, M, I)^{\beta_k}$ est une fonction totalement plate (égale à 1) et dépourvue de maxima : Les mouvements dans un tel espace sont totalement libres et d'un point de vue physique cela signifie que l'énergie nécessaire pour se déplacer d'un point A à un point B de l'espace des paramètres est nulle (il n'y a plus de barrière de potentiel). La loi *a posteriori* $\pi_k(\boldsymbol{\theta}^{\boldsymbol{k}}_{(\boldsymbol{i})}|y, M, I)$ est alors égale à l'*a priori*.

Dans la pratique, le nombre de chaines de Markov parallèles est discret et il est légitime de se demander quel est le meilleur nombre m et la meilleure distribution en température $T = \{T_1, T_2, ..., T_m\}$ à adopter. En effet même si le *parrallel tempering* se prête facilement au calcul parallèle (sur plusieurs CPU), le temps de calcul sur un processeur est proportionnel au nombre de chaines utilisées.

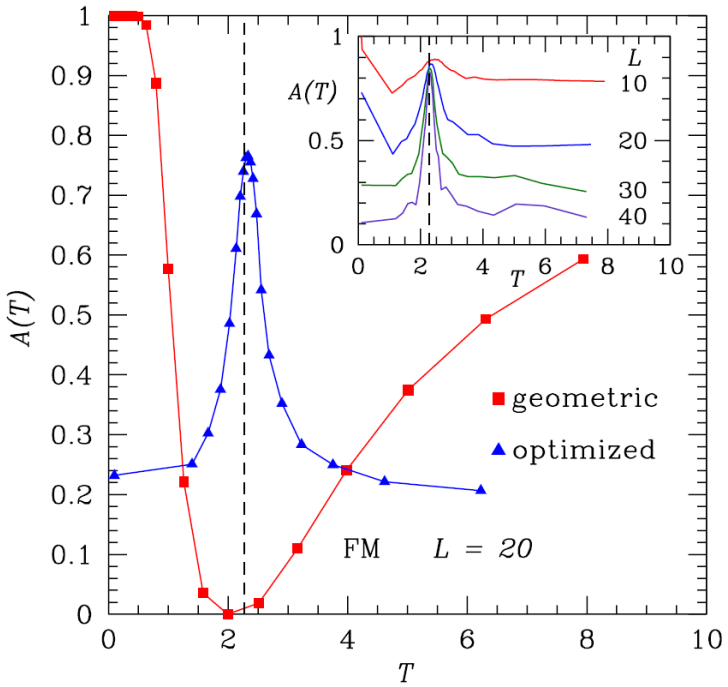


FIGURE 3.12 – Un des résultats obtenu dans l'article de Katzgraber (2006) pour un modèle ferromagnétique d'Ising. On y montre les conséquences du choix d'une loi géométrique pour la loi de température. Il est évident que dans ce genre de situation, le choix de loi géométrique n'est pas approprié puisqu'à la température critique le taux de permutation $A(T)$ (l'équivalent de notre α_{switch}) de chaine tombe à 0 : la loi géométrique n'est pas efficace à la température critique de changement d'état. L'encart montre le taux de permutation en fonction de la température pour différente valeur de la taille L de la boite de dipôles, dans le cas optimisé.

Les approches plus conventionnelles supposent que la loi en température $T = f(k)\ \forall k \in [1, m]$ est obtenue lorsque la probabilité de transition entre deux chaines de Markov tempérées k et $k+1$ est constante. Alors, une approximation raisonnable de $T = f(k)$ est une loi géométrique Cristian et al. (2004) du type $T_k = \lambda^k$. Cependant, si il existe une transition de phase violente tel que

dans les modèles (ferromagnétiques) d'Ising, cette loi est caduque car il existera alors un *goulot d'étranglement* (*bottleneck*) à la température de transition de phase (cf. Fig.3.11 et Fig.3.12 pour une illustration dans le cas du modèle d'Ising). Néanmoins nous pensons que dans le cadre des applications que nous faisons des TMCMC, les transitions de la sorte ne sont pas significatives : les multimodalités de la distributions d'intérêts ne sont pas trop prononcées ni trop nombreuses.

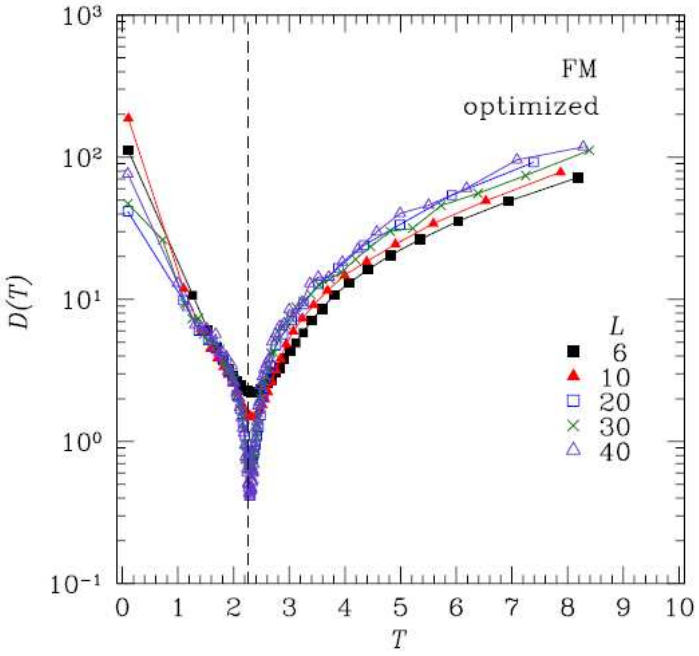


FIGURE 3.13 – Un des résultats obtenu dans l'article de Katzgraber (2006) pour un modèle ferromagnétique d'Ising. On y montre la diffusivité local $D(T)$ du processus d'exploration de l'espace par des chaines de Markov tempérées en fonction de la température T (ici T est une grandeur pouvant être interprétée physiquement), après optimisation par son algorithme d'ajustement automatique de la distribution des températures. Les différents symboles et couleurs représentent le nombre de dipôles par coté d'une boite pour un modèle d'Ising 2D. Il y L^2 dipôles par boite. On voit clairement que dans un tel modèle physique, il existe une température critique à partir de laquelle le réseau de dipôles change brutalement de structure : On passe d'une phase d'ordre magnétique à basse température (tous les dipôles orientés dans le même sens et le milieu est magnétique) à une phase paramagnétique (les dipôles sont orientés dans des directions aléatoire et la magnétisation du milieu est perdu en dehors de l'influence d'un champ magnétique extérieur).

3.2.5 Comparaison de modèles avec un algorithme MCMC tempéré

Un autre des avantages majeur du *parrallel tempering* est de permettre théoriquement le calcul exact de la probabilité d'un modèle. En effet, comme montré par Gregory (2005b), une approche Bayésienne complète nécessite le calcul du *facteur de Bayes* B_{M_1,M_2} pour déterminer le coefficient concurrentiel,

$$B_{M_1,M_2} = \frac{P(y|M_1,I)}{P(y|M_2,I)}, \tag{3.41}$$

où $P(y|M_1, I)$ et $P(y|M_2, I)$ sont les *evidence* (ou *global likelihood*) de chacun des modèles. En effet, *via* la règle de Bayes, le coefficient concurrentielle s'écrit,

$$O_{M_1,M_2} = \frac{P(M_1|y, I)}{P(M_2|y, I)} = \frac{P(M_1|I)}{P(M_2|I)} B_{M_1,M_2}. \tag{3.42}$$

En règle générale, nous n'avons pas de connaissance *a priori* quantitative sur chacun des modèles. Dans ce cas, on impose naturellement $P(M_1|I) = P(M_2|I) = 1/2$.

La comparaison de modèle dans un espace des modèles discret de taille N_{mod} s'écrit,

$$P_R(M_j|y, I) = \frac{P(y|M_j, I)P(M_j|I)}{\sum_{i=1}^{N_{mod}} P(y|M_i, I)P(M_i|I)} \tag{3.43}$$

$$= \frac{1}{1 + \sum_{i=1}^{N_{mod}} O^{-1}_{M_j,M_i}} \tag{3.44}$$

Cette équation est simplement une généralisation de l'Eq.2.10.

Les termes d'*evidence* sont obtenus par intégration sur tous les paramètres de la *loi jointe*,

$$P(y|M_i, I) = \int \psi(\boldsymbol{\theta}, y|M_i, I)d\boldsymbol{\theta} = \int \pi(y|\boldsymbol{\theta}M_i, I)\pi(\boldsymbol{\theta}|M_i, I)d\boldsymbol{\theta}. \tag{3.45}$$

Cette opération est dénommé *marginalisation*. Sous cette forme, l'intégrale n'est pas utilisable en pratique car nous ne disposons que du logarithme des probabilités (*log-probabilités*) connus à une constante près. Dans la suite, nous allons la modifier pour faire apparaitre explicitement le paramètre $\beta = 1/T$ et montrer que le calcul de la log-probabilité $ln\, P(y|M_i, I)$ n'est rien d'autre qu'un calcul de moyenne sur β de log-quantités.

Soit une fonction de partition $Z(\beta)$ correspondant à l'intégrale de la *loi jointe* de chacune des chaines de Markov parallèle,

$$Z(\beta) = \int \pi(\boldsymbol{\theta}|M_i, I)\pi(y|\boldsymbol{\theta}, M_i, I)^\beta d\boldsymbol{\theta}$$

$$= \int e^{ln\,\pi(\boldsymbol{\theta}|M_i,I)+\beta ln\,\pi(y|\boldsymbol{\theta},M_i,I)} d\boldsymbol{\theta} \tag{3.46}$$

La dérivée de cette fonction s'écrit,

$$\frac{d}{d\beta}Z(\beta) = \int \frac{d}{d\beta} e^{ln\,\pi(\boldsymbol{\theta}|M_i,I)+\beta ln\,\pi(y|\boldsymbol{\theta},M_i,I)} d\boldsymbol{\theta}$$

$$= \int \pi(\boldsymbol{\theta}|M_i, I) ln\,\pi(y|\boldsymbol{\theta}, M_i, I)\pi(y|\boldsymbol{\theta}, M_i, I)^\beta d\boldsymbol{\theta}. \tag{3.47}$$

En utilisant la relation $\frac{d}{d\beta} ln\, Z(\beta) = \frac{1}{Z(\beta)} \frac{d}{d\beta} Z(\beta)$ et en la combinant avec l'équation précédente (Eq.3.47) on obtient,

$$\frac{d}{d\beta} ln\, Z(\beta) = \frac{\int ln\,\pi(y|\boldsymbol{\theta}, M_i, I)\pi(\boldsymbol{\theta}|M_i, I)\pi(y|\boldsymbol{\theta}, M_i, I)^\beta d\boldsymbol{\theta}}{\int \pi(\boldsymbol{\theta}|M_i, I)\pi(y|\boldsymbol{\theta}, M_i, I)^\beta d\boldsymbol{\theta}}. \tag{3.48}$$

On remarque que cette fonction est de la forme $\int g(\boldsymbol{\theta}) f(\boldsymbol{\theta}, \beta) d\boldsymbol{\theta} / \int f(\boldsymbol{\theta}, \beta) d\boldsymbol{\theta}$, ce qui correspond à la définition d'une moyenne de $g(\boldsymbol{\theta}) \equiv ln\,\pi(y|\boldsymbol{\theta}, M_i, I)$, pondérée par une fonction de β. Par la suite, on appellera $< ln\,\pi(y|\boldsymbol{\theta}, M_i, I) >_{\boldsymbol{\theta}|\beta}$ (moyenne selon $\boldsymbol{\theta}$ et à β donné) cette moyenne pondérée.

On fera remarquer que dans la limite où $N \to \infty$, ce terme correspond à une estimée de $ln\,\pi(y|\boldsymbol{\theta}, M_i, I)$ et ce indépendamment de β. C'est une autre manière de dire que tous les états possibles de l'espace des paramètres ont été visités au bout d'un temps infini (propriété d'ergodicité garantie) et ce, quelque soit la température utilisée.

Nous ne disposons malheureusement pas d'un nombre d'échantillons infini et il nous faut supprimer toute dépendance à la température et utiliser pleinement tous les échantillons que nous avons à notre disposition. Pour cela on somme selon β la fonction $< ln\,\pi(y|\boldsymbol{\theta}, M_i, I) >_{\boldsymbol{\theta}|\beta}$, ce qui nous amène à ré-écrire le terme de gauche de l'Eq.3.48 sous forme intégrée,

$$\begin{aligned} \int_0^1 dln\,Z(\beta) &= ln\,Z(\beta=1) - ln\,Z(\beta=0) \\ &= \int < ln\,\pi(y|\boldsymbol{\theta}, M_i, I) >_{\boldsymbol{\theta}|\beta} .d\beta. \end{aligned} \tag{3.49}$$

Les bornes de l'intégrale $\beta = 1$ et $\beta = 0$ sont déduites directement des conditions aux limites et de la relation $\beta = 1/T$. La *distribution d'intérêt* correspond à $\beta = 1$ et $\beta = 0$ correspond à la chaine tempérée $T \to \infty$.

Il ne nous reste plus qu'à expliciter les termes $Z(\beta = 0)$ et $Z(\beta = 1)$ pour aboutir. En reprenant l'Eq.3.46 on a simplement,

$$Z(\beta = 0) = \int \pi(\boldsymbol{\theta}|M_i, I)d\boldsymbol{\theta} \tag{3.50}$$

$$Z(\beta = 1) = \int \pi(\boldsymbol{\theta}|M_i, I)\pi(y|\boldsymbol{\theta}, M_i, I)d\boldsymbol{\theta} \equiv \pi(y, |M_i, I). \tag{3.51}$$

On remarquera que $Z(\beta = 1)$ correspond au terme intégral que nous cherchons à calculer. $Z(\beta = 0)$ correspond à l'intégrale de l'*a priori*. On retrouve donc ce que nous disions de manière qualitative quelques pages précédentes lorsque nous introduisions la notion de *parrallel tempering* : à haute température, l'*a priori* domine vis à vis de la vraisemblance tempérée.

A condition que les *a priori* soient normalisés (et normalisables), alors $Z(\beta = 0) = 1$ et on aboutit à une relation relativement simple,

$$\begin{aligned} ln\,\pi(y|M_i, I) &= ln\,Z(\beta=0) + \int < ln\,\pi(y|\boldsymbol{\theta}, M_i, I) >_{\boldsymbol{\theta}|\beta} \\ &= \int < ln\,\pi(y|\boldsymbol{\theta}, M_i, I) >_{\boldsymbol{\theta}|\beta} \end{aligned} \tag{3.52}$$

On traite avec des grandeurs discrètes. Réécrivons les grandeurs d'intérêts de manière appropriée :

$$< ln\,\pi(y|\boldsymbol{\theta}, M_i, I) >_{\boldsymbol{\theta}|\beta} \approxeq \frac{1}{N}\sum_t ln\,\pi(y|\boldsymbol{\theta}_{\boldsymbol{t}}^{(\boldsymbol{\beta})}, M_i, I). \tag{3.53}$$

$$ln\,\pi(y|M_i, I) \approxeq \sum_k < ln\,\pi(y|\boldsymbol{\theta}, M_i, I) >_{\boldsymbol{\theta}|\beta_k} . \tag{3.54}$$

Avec ces deux dernières relations, nous voyons que la précision du résultat de l'estimation dépendra à la fois du nombre d'échantillons et du nombre de chaines parallèles. Pour obtenir une précision relativement bonne (mais non quantifiée (!)) sur l'Eq.3.54, Gregory (2005b) propose d'effectuer une interpolation quadratique de $ln\,\pi(y|M_i, I)$. Le gros défaut de cette approche est d'être dans l'incapacité de fournir une barre d'erreur sur l'estimé des probabilités. Lorsque nous utilisons les Eq.3.53 & 3.54, il faut être relativement prudent et s'assurer que :

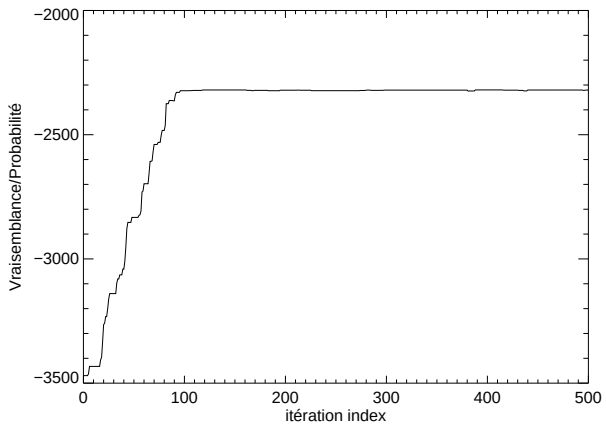


FIGURE 3.14 – Exemple de moyen de diagnostique de la chaine de Markov : Lorsque l'état stationnaire est atteint, la moyenne flottante de la vraisemblance/probabilité est stable dans le temps. La phase de chauffage se situe entre 0 et 100 itérations ici. Les fluctuations autour de l'état stationaire ne sont pas visible sur ce graphe pour des raisons d'échelles.

1/ La distribution en température est suffisamment représentative de la fonction $< ln\,\pi(y|\boldsymbol{\theta}, M_i, I) >_{\boldsymbol{\theta}|\beta_k}$. Autrement, on peut être confronté aux problèmes de goulot d'étranglement exposé au Chap.3.2.4.

2/ La *distribution stationnaire* est atteinte pour chacune des distributions *a posteriori* échantillonnée (*i.e.* à toutes les températures β).

3/ Le nombre d'échantillon est représentatif pour chacune des distributions *a posteriori* échantillonnés (ce qui présuppose que la condition (2) soit remplie).

Sur ce point nous arrivons aux limites de l'algorithme de Metropolis-Hasting et du *parralel tempering*. Certifier que la distribution stationnaire est atteinte est relativement délicat, voire impossible comme l'ont souligné Brooks & Roberts (1998) ou encore Cowles & Carlin (1996). Certains outils de diagnostiques parfois relativement complexes permettent de s'assurer que l'échantillonnage est représentatif. Dans notre cas, pour s'assurer que la distribution stationnaire était atteinte, nous nous sommes basés sur les graphes présentant la log-probabilité en fonction des index d'échantillons (cf. Fig.3.14) : c'est l'approche la plus simple et intuitive. Enfin pour s'assurer que le nombre d'échantillon était relativement suffisant, nous divisions en deux part égales la série d'échantillons acquis au préalable et calculions pour chacune de ces sous-séries la probabilité du modèle. Si les deux valeurs variaient de manière significative comparativement à la valeur moyenne du facteur de Bayes, alors on doublait le nombre d'échantillons en lançant une acquisition supplémentaire. Pour fixer les idées prenons un exemple. Si nous avons facteur de Bayes moyen de 10, avec des variations de 8 à 12, il n'y a pas lieu de doubler le nombre d'échantillons. Ce n'est plus le cas si nous avons un facteur de Bayes de 4 et des variations de 0.5 à 12 : On est plus capable de trancher car le facteur de Bayes privilégie l'un ou l'autre des modèles.

Une alternative possible serait d'utiliser des algorithmes MCMC dit *coupling from the past* tel que celui de Propp & Wilson (1996). Ces algorithmes, basés sur un schéma numérique radicalement différent de celui de Metropolis, sont capables d'estimer le nombre d'échantillons nécessaires pour que la fonction à échantillonner le soit correctement. Une autre alternative intéressante consisterait à utiliser le *Nested sampling* (échantillonnage en grappes). C'est une

technique proposée en 2004 qui consiste à sélectionner automatiquement les secteurs où la distribution d'intérêt est significative et à n'échantillonner que ces zones [11]. Cet outil a été utilisé en cosmologie (problème de faibles dimensions, *i.e.* $N \lessapprox 20$) avec succès et semble prometteuse car elle est plus rapide que les MCMC classiques et peut à la fois calculer la probabilité d'un modèle et son incertitude. On peut aussi citer comme alternative, les réseaux de neurones dont les performances en terme de sélection de modèle dans des problèmes complexes (très grandes dimensionnalités *i.e.* $N >> 100$) ne sont plus à prouver.

3.2.6 Résumé sur la méthode d'échantillonnage

Pour clore ce chapitre et comme un bon schéma vaut bien plus que de beaux discours, nous résumons sur la Fig.3.15 les éléments de base de l'algorithme *Tempered and Adaptive MCMC* (TAMCMC) que nous avons développée. On remarquera que lors de la phase d'apprentissage, le mélange de chaines tempérées n'est pas opéré. Ce choix résulte de test que nous avons menés : mélanger les chaines parallèles durant le processus d'apprentissage semble perturber la distribution d'équilibre de la chaine de Markov et celui ci n'est plus représentatif de la distribution d'intérêt. En conséquence lors de l'échantillonnage proprement dit, le taux d'acceptation ne correspond pas au taux τ que l'on cherchait à atteindre durant l'apprentissage.

11. se référer à http ://www.inference.phy.cam.ac.uk/bayesys/. Ce site référence et propose des programmes libres de Nested sampling.

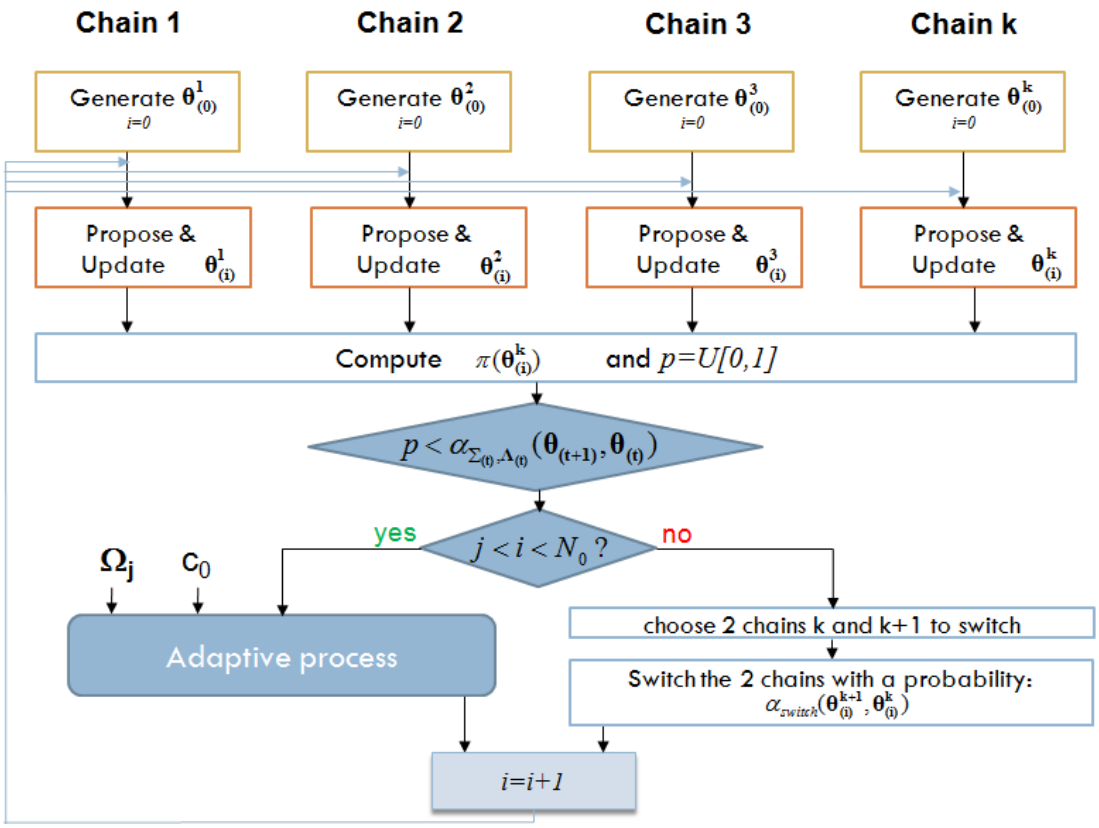


FIGURE 3.15 – Représentation schématique de mon algorithme. Il est constitué d'un algorithme de Metropolis, couplé à un algorithme adaptatif et mettant en oeuvre le *parallel tempering*. N_0 correspond à l'index itératif maximum pour lequel l'apprentissage est effectué. Chaque chaine parallèle dispose d'une matrice de covariance indépendante ajustée par le processus adaptatif.

Chapitre 4

Méthodes, simulations et applications des MCMC sur des étoiles de CoRoT

Sommaire

4.1 Quel méthodologie adopter ? 71
4.1.1 Une stratégie globale d'ajustement du spectre de puissance 71
4.1.2 Une approche bayésienne pour les pulsateurs de type solaire 73
4.2 Simulations 76
4.2.1 Fonction de densité de probabilité et choix de critères statistiques . . . 76
4.2.2 Spectres stellaire simulés 86
4.3 Applications aux étoiles 91
4.3.1 Un cas stellaire intensivement étudié : HD49933 91
4.3.2 Un autre cas difficile : HD181420 100

Dans ce chapitre, nous allons présenter les possibilités offertes par une analyse Bayésienne avec l'algorithme que nous avons développé (cf. chapitre précédent). Avant de commencer notons que des analyses Bayésienne par échantillonnages MCMC de pulsateurs solaire ont déjà été effectuées par le passé par Brewer et al. (2007), mais celles ci reposaient sur des hypothèses incorrect concernant la nature de l'excitation des modes (et donc aussi sur la statistique du bruit). Nous commencerons par présenter notre vision de l'analyse bayésienne appliquée aux pulsateurs de type solaires. Puis nous présenterons quelques exemples d'analyses en nous basant sur des spectres simulés. Enfin nous présenterons des analyses d'étoiles observées par CoRoT et survolerons les résultats obtenus avec Kepler jusqu'à ce jour.

4.1 Quel méthodologie adopter ?

4.1.1 Une stratégie globale d'ajustement du spectre de puissance

Les spectres de puissance calculés à partir des séries temporelles portent à la fois la signature du bruit instrumental, stellaire (granulation à différentes échelles de temps caractéristique) et de chacun des modes de pulsation. De plus, notre connaissance *a priori* sur le signal est relativement faible en général. De notre point de vue il faut donc,

1/ identifier les propriétés globales du bruit de par le spectre.

2/ identifier les propriétés globales des modes de pulsation à savoir, le domaine de fréquence d'excitation des modes.

3/ lorsque la loi asymptotique est valide, estimer ses valeurs caractéristiques.

4/ établir l'ajustement des modes individuels sur la base des connaissances sur le signal obtenus grâce aux étapes 1, 2 et 3.

C'est avec cette logique en tête que toutes nos analyses ont été menées. Pour le cas des étoiles CoRoT, celles ci étant en nombre restreint, des observations visuelles préliminaires permettent de définir *a priori* certaines propriétés, tel que par exemple le nombre de modes à ajuster dans le spectre et en conséquence, la plage de fréquence sur laquelle les modes sont supposés significatifs vis à vis du bruit (ce n'est évidement plus le cas si nous souhaitons établir une procédure automatique. Une approche possible est discuté Chap.5 dans ce cas). Un corolaire de cette dernière affirmation est qu'il est inutile *a priori* d'établir un ajustement de tout le spectre à partir du moment où nous ne sommes intéressés que par les modes de pulsation. A contrario, ne pas inclure une plage suffisamment large de fréquence lors de l'ajustement peut significativement nuire à la qualité de la détermination des paramètres associés au bruit et aux modes car comme présenté au Chap.1, le bruit stellaire est une fonction de la fréquence. Une estimation fiable de ce dernier nécessite une étendue spectrale importante. Or la durée de calcul d'un modèle du spectre de puissance augmente comme $Nlog(N)$, N étant le nombre de points considérés dans l'intervalle spectrale soumis à ajustement [1]. On se retrouve donc face à un dilemme.

L'approche Bayésienne de part sa flexibilité permet de concilier les deux points de la discussion précédente, apparemment contradictoires. En effet, une solution consiste à établir une chaines d'ajustements successifs caractérisée par des niveaux descriptifs différents. Il est alors possible de se servir de l'information collectée à une échelle globale et de s'en servir en tant qu'*a priori* pour un ajustement à une échelle plus réduite. Dans les faits, et en se basant sur ce qui a été écrit ci dessus on utilise une approche en 3 étapes,

Etape 1 (E1)/ Le spectre de puissance est ajusté sur une large plage de fréquence, typiquement de 0 à $F_c/2$ (fréquence de Nyquist) avec un modèle grossier des modes et une description la plus fine possible du bruit. Le bruit est alors décrit par 3 semi-lorentziennes (*Harvey-like* en anglais ou profil de *type Harvey*) et un bruit blanc. L'enveloppe des modes est décrite par une fonction gaussienne.

Etape 2 (E2)/ On estime la grande séparation $\Delta\nu$ et la valeur de ϵ (loi asymptotique). Pour $\Delta\nu$, cela peut se faire par autocorrélation du spectre de puissance. On obtient ainsi la périodicité moyenne du motif. ϵ peut être évalué soit visuellement, soit par exemple, par un ajustement d'un motif de fréquence contraint par la loi asymptotique.

Etape 3 (E3)/ On restreint la zone d'ajustement à la plage avérée de présence des modes. Pour améliorer la stabilité de l'ajustement, une plage d'environ 1 mHz est ajoutée à haute fréquence. Elle est supposée ne contenir que du bruit. Le nombre de semi-lorentziennes jugé suffisant pour décrire le bruit est évalué soit statistiquement par comparaison de modèles (cela peut se révéler couteux en temps de calcul), soit défini sur la base d'arguments qualitatifs. Dans ce dernier cas, cela consiste en une observation visuelle de chacun des profils de Harvey : on identifie ainsi les profils dont l'influence est négligeable dans la zone d'excitation des modes. Les modes sont décrits par un modèle global tenant compte de la rotation, des fréquences, largeurs et hauteurs individuelles. Les estimés de $\Delta\nu$ et ϵ permettent de définir un vecteur initial de paramètres pour les fréquences. Ces derniers permettent aussi d'établir un *a priori* limitant le volume d'espace dans lequel l'échantillonnage s'effectue.

1. Le spectre de puissance en lui même est calculé une seule fois. Son temps de calcul est donc négligeable vis à du temps de calcul du modèle à ajuster.

Paramètre	Nature de l'*a priori*
Paramètres des modes	
$H_{l=0}$ $(ppm^2/\mu Hz)$	$J(1,10)$
$V_{l=1}$	$J(1.5,5)$
$V_{l=2}$	$J(0.5,5)$
Γ (μHz)	$J(5,50)$
Angle (°)	$U[0,90]$
Paramètres de bruit	
A_i $(ppm^2/\mu Hz)$	$N(< A_i >_{E1}, \sigma_{A_{iE1}})$
B_i (μHz)	$N(< B_i >_{E1}, \sigma_{B_{iE1}})$
p_i (pas d'unité)	$N(< p_i >_{E1}, \sigma_{p_{iE1}})$
C $(ppm^2/\mu Hz)$	$N(< C >_{E1}, \sigma_{C_{E1}})$

TABLE 4.1 – Nature et valeurs caractéristiques des *a priori* utilisés, communs aux étoiles CoRoT analysées. Pour les étoiles Kepler, quelques ajustements ont été effectués sur la hauteur tolérée de $H_{l=0}$ (borne supérieure relevée). Pour la signification des sigles désignant les *a priori*, se référer à l'Annexe A. les valeurs au format $\langle x \rangle_E$ et $\sigma_{E_{iE}}$ désignent respectivement la valeur moyenne et l'écart type du paramètre obtenu lors de l'étape E.

4.1.2 Une approche bayésienne pour les pulsateurs de type solaire

Maintenant que la logique globale adopté a été présenté, nous allons présenter de manière plus formelle les *a priori* communs utilisés lors des diverses analyses menées pour les hauteurs, largeurs, les paramètres du bruit et l'angle d'inclinaison. On retrouvera leurs valeurs types dans les cas stellaires étudiés dans la Table 4.1. Les principes du choix des *a priori* sur le splitting et les fréquences sont aussi abordés ici et leurs valeurs caractéristiques dans les cas stellaires étudiés sont présenté Table 4.2.

Comme déjà souligné au Chap.1, le spectre de puissance d'une étoile de type solaire est constitué en première approximation d'un peigne de fréquence défini par la loi asymptotique. Chacune de ces fréquences est caractérisée par 3 nombres quantique (n, l, m). Nous nous servons de cette loi pour définir les *a priori* fréquentiels pour chacune des identifications possibles dont les modèles associés seront appelés M_A et M_B dans la suite. Rappelons la loi asymptotique,

$$\nu(n,l) \simeq (n + \epsilon + l/2)\Delta\nu - l(l+1)D_0. \tag{4.1}$$

On en déduit la valeur moyenne des petites séparations pour les degrés les plus bas,

$$\delta\nu_{02}(n) = \nu_{n+1,2} - \nu_{n,0} \simeq 6D_0 \tag{4.2a}$$

$$\delta\nu_{13}(n) = \nu_{n+1,3} - \nu_{n,1} \simeq 10D_0. \tag{4.2b}$$

Discriminer entre M_A et M_B revient à imposer un *a priori* sur le paramètre $\Delta\nu$ mais surtout sur ϵ. Lorsque $\Delta\nu$ est connue de manière précise, ajouter ou soustraire 1/2 à ϵ revient en effet à inverser les deux identifications. D_0 ne joue pas un rôle crucial sur l'identification des couples pair/impair de modes de bas degré car $D_0 << \Delta\nu$. C'est le rôle de l'étape 2 (E2) que d'obtenir de tels *a priori*. Cependant, il faut aussi évaluer le degré de tolérance autour de la position moyenne fournie par E2.

La plupart du temps nous avons choisi de contraindre les fréquences par un *a priori* $\pi(\nu_{n,0}, \nu_{n,1}|M_j, I)$ gaussien et en conséquence, la tolérance correspond à son écart type σ. Pour

Indicatif de l'étoile	Paramètre	Nature de l'*a priori*
HD 49933 (IRa-01)	ν_s (μHz)	$N(3.4, 0.34)$
	$\Delta\nu$ (μHz)	$N(85.6, 0.8)$
	ϵ (pas d'unité)	1.55 (M_A) / 1.05 (M_B)
	D_0 (μHz)	$D_0 = 1$ (Fixe)
	$\delta\nu_{02}$ (μHz)	$N(6, 3)$
HD 49933 (IRa-01+LRa-01)	ν_s (μHz)	$N(3.4, 0.34)$ ou $N(3.4, 0.68)$
	$\Delta\nu$ (μHz)	$N(85.7, 0.6)$
	ϵ (pas d'unité)	1.52 (M_A) / 1.02 (M_B)
	D_0 (μHz)	$D_0 = 1$ (Fixe)
	$\delta\nu_{02}$ (μHz)	$N(6, 3)$ ou $N(6, 6)$
HD 181420 (LRc-01)	ν_s (μHz)	$N(4.4, 1.7)$
	$\Delta\nu$ (μHz)	$N(75.3, 1.0)$
	ϵ (pas d'unité)	1.4 (M_A) / 1.9 (M_B)
	D_0 (μHz)	$D_0 = 1$ (Fixe)
	$\delta\nu_{02}$ (μHz)	$N(6, 3)$

TABLE 4.2 – Tableau récapitulatif des *a priori* de nature informative utilisés dans quelques cas stellaires étudiés. La définition des sigles est donnée en Annexe A.

le $l = 0$ et le $l = 1$, nous définissons,

$$\pi(\nu_{n,0}, \nu_{n,1}|M_j, I) = \frac{1}{2(n_{max} - n_{min})\sigma\sqrt{2\pi}} \sum_{n=n_{min}}^{n_{max}} [\mathrm{e}^{-\frac{(\nu-\nu(n,0))^2}{2\sigma^2}} + \mathrm{e}^{-\frac{(\nu-\nu(n,1))^2}{2\sigma^2}}]. \tag{4.3}$$

Dans cette équation, $n_{max} - n_{min}$ représente le nombre d'ordre radiaux tenus en compte lors de l'ajustement.

Pour ce qui concerne la fréquence du $l = 2$, les contraintes diffèrent en fonction de l'étoile considérée. Dans les cas où le signal sur bruit est insufisant (HD49933 ou HD181420), ce mode n'est peu ou pas visible et on utilise des contraintes gaussiennes,

$$\pi(\delta\nu_{02}(n)|M_j, I) = \frac{1}{\sigma_{02}\sqrt{2\pi}} \mathrm{e}^{-\frac{(\delta\nu_{02}(n) - <\delta\nu_{02}>_{th})^2}{2\sigma_{02}^2}}. \tag{4.4}$$

La valeur attendue de la petite séparation est nommée $< \delta\nu_{02} >_{th}$. On tolère des variations de l'ordre de 50% à 100% de $< \delta\nu_{02} >_{th}$ en fonction des modèles testés. Au final, la contrainte appliquée sur le mode est relativement forte : on tente de compenser le manque d'information dans le spectre par l'apport d'une information théorique.

A l'opposé, lorsque les données sont jugées de qualité comme pour HD52265 [2], on laisse parler les données et on utilise un *a priori* moins informatif, de type gaussien-uniforme-gaussien (GUG, cf. Annexe A). L'*a priori* est défini de tel sorte que l'on évite la superposition des modes $l = 0$ et $l = 2$, sans toutefois l'interdire (cf. Fig.4.1).

Du fait de l'atmosphère stellaire et de l'effet centre-bord le nombre de modes de degré l visibles est limité. Ainsi dans le cas solaire, si la hauteur relative moyenne (vis à vis du $l = 0$) des modes de degré $l = 1$ est de 1.5, elle n'est plus que de 0.53 pour $l = 2$ et de 0.027 pour les

2. Cette étoile CoRoT n'a pas encore fait l'objet d'une publication. En accord avec les règles de publications nous ne montrerons donc pas les résultats obtenues sur cette étoile.

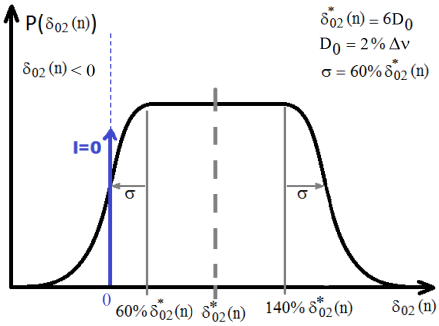


FIGURE 4.1 – A priori "gaussien-uniforme-gaussien" (GUG) appliqué à la petite séparation $\delta_{02}(n)$ utilisé lorsque le rapport hauteur sur bruit est bon et que le $l = 2$ est facilement identifiable.

$l = 3$ du fait de l'observation non imagée de l'étoile. Les degrés supérieurs ont tous des hauteurs relatives égales ou inférieures à celle du $l = 3$.

En asterosismologie, le rapport signal à bruit est très souvent inférieur à 10. Dans ces conditions, on ne peut envisager sérieusement un ajustement des modes de degré $l = 3$ ou plus pour des étoiles dont les processus physiques (la source d'excitation des modes par exemple) sont similaires à ceux en oeuvre dans le soleil. Rappelons que nous soupçonnons l'atmosphère et l'effet centre-bord de pouvoir modifier significativement ces valeurs dans certains cas. L'examen des étoiles de Kepler semble par exemple montrer que les étoiles de type solaire les plus évoluées (ayant quitté la séquence principale) ont des hauteurs relatives pour le $l = 2$ qui peuvent être très supérieures à 0.53 (allant jusqu'à 1). Il est donc possible que le $l = 3$ soit visible pour certaines de ces étoiles. Ceci est cohérent avec la découverte de 2 modes identifiés comme des $l = 3$ dans une étoile CoRoT analysée par Deheuvels et al. (2010) : HD 49385.

Pour les différentes étoiles présentées dans ce manuscrit, nous avons souvent établi des ajustements avec des *a priori* différents sur la hauteur des modes. Nous n'avons pas inclu le $l = 3$ car des tests préliminaires à l'ajustement n'ont pas permis de les mettre en évidence (collapsogrammes, diagramme échelle). Ces étoiles ne sont pas *a priori* très évoluées. Les situations envisagées sont listées ci-dessous.

1/ $V_{l=1} = \frac{H_{l=1}}{H_{l=0}}$ et $V_{l=2} = \frac{H_{l=2}}{H_{l=0}}$ libres, suivant un *a priori* de Jeffrey.

2/ $V_{l=1}$ libre suivant un *a priori* de Jeffrey. $V_{l=2}$ fixé à la valeur solaire.

3/ $V_{l=1}$ et $V_{l=2}$ fixés à la valeur solaire.

Les largeurs sont aussi contraintes avec un a priori de Jeffrey. Ce type d'*a priori* est non informatif et uniforme dans le domaine logarithmique. Pour sa formulation mathématique et un schéma, se référer à l'Annexe A. Jeffrey (1961) a justifié ce choix en spécifiant qu'il est invariant en translation pour des paramètres de nature intensive. Par opposition, l'a priori non-informatif uniforme est plutôt recommandé pour des paramètres de position (comme la fréquence par exemple). On soulignera que dans les faits, aucun *a priori* est non-informatif : en fait, tout *a priori* purement non informatif est impropre (l'intégrale diverge). Imposer un *a priori* de Jeffrey semble toutefois permettre une meilleure discrimination qu'un *a priori* uniforme. C'est par ce que l'une de ces valeurs caractéristiques permet de trancher plus franchement entre bruit et signal. En effet, la densité de probabilité de cette fonction est en $1/x$ pour x proche de 0, mais est plutôt uniforme si $x > x_{min}$, x_{min} étant un paramètre de l'*a priori* définissant la zone de transition supposée entre signal et bruit. On suppose aussi que les variations de largeurs ne sont pas brutales en fonction de la fréquence de telle sorte que nous nous autorisons à les supposer comme une fonction croissante de n (indépendante de l et m). C'est donc une fonction de la

fréquence, croissante par plateau.

Pour le splitting rotationnel, on suppose qu'il est constant sur toute la plage des modes ajustés. Cette hypothèse est essentiellement justifiée par le faible rapport signal à bruit auquel on fait face. Le spectre ne contient *a priori* pas assez d'information pour permettre une description plus fine. L'*a priori* se base sur la présence/absence de pics caractéristiques à très basse fréquence dûs à la rotation de l'étoile. La présence de structures à la surface de l'étoile dont le temps de vie est supérieur à la période de rotation P_{rot} engendre une modulation périodique de l'intensité lumineuse. Dans le spectre de puissance, plusieurs pics d'intensités apparaissent, à $1/P_{rot}$ et à ses harmoniques.

Pour les étoiles HD49933 et HD181420, ces pics sont clairement visibles et on estime leur rotations respectivement à 3.4 jours et 2.8 jours. Pour ces étoiles, on utilise un *a priori* gaussien. Pour HD52265, et les étoiles Kepler étudiées, la rotation est moins évidente à évaluer. On suppose qu'elle n'est pas très rapide : l'*a priori* est uniforme pour des périodes de rotation supérieures à 5.8 jours ($\nu_s = 2$ μHz). Les rotations plus rapides sont considérées comme peu probables et contraintes par une gaussienne. C'est pourquoi cet *a priori* est nommé uniforme-gaussien (UG, cf. Annexe A).

Les contraintes à imposer sur le bruit lors de l'ajustement des paramètres individuels des modes sont cruciales car les modes sont souvent noyés dans le bruit. L'ajustement de l'étape E1 permet d'établir des contraintes fiables sur les différentes fonctions de type Harvey. Durant cette étape, nous effectuons un ajustement mettant en oeuvre une méthode par MLE (absence d'a priori). Nous imposons ensuite un *a priori* gaussien dont la valeur centrale et l'écart type correspondent au résultat obtenu par MLE.

4.2 Simulations

4.2.1 Fonction de densité de probabilité et choix de critères statistiques

Dans un premier temps, pour tester les conditions de fonctionnement du code que nous avons élaboré, nous avons effectué une série de simulations. L'idée principale était d'évaluer les capacités de la comparaison de modèles et de définir les critères statistiques les plus pertinents à utiliser. Pour cela, nous avons construit des spectres synthétiques répondant à la statistique du χ^2_2 ne contenant que 3 modes : $l = 0$, $l = 1$, $l = 2$. De fait, nous nous placions dans une situation où on simule un ajustement de modes de pulsation stellaire présent seulement sur une grande séparation. Si on suppose le bruit, la largeur des modes et le splitting rotationnel localement constant, un tel modèle ne nécessite que 10 paramètres : 3 fréquences, 3 hauteurs, 1 largeur, 1 splitting rotationnel, 1 angle d'inclinaison stellaire et 1 bruit blanc.

La motivation initiale de l'utilisation des TAMCMC était de savoir si, avec une approche bayésienne mettant en oeuvre une récupération complète de l'information portée par le critère statistique, nous pouvions, en théorie du moins, identifier les degrés des modes pairs et impairs dans des situations similaires à la première étoile observée par CoRoT : HD49933. Lorsque ces simulations ont été menées, nous savions déjà que les conditions d'analyses des données CoRoT dans le cas de cet étoile était particulièrement difficiles (grâce aux analyses effectués avec des approches par maximum de vraisemblance). Cela est vraisemblablement dû à plusieurs effets cumulés : une largeur des modes bien plus importante que dans le cas solaire, un splitting rotationnel voisin de la petite séparation $\delta\nu_{02}$ et un faible rapport entre hauteurs des modes et bruit local ($H/N < 7$). Des modèles préliminaires de l'étoile basés sur les informations issues de la campagne de mesure dirigée par Mosser et al. (2005) avait déjà permis d'obtenir une estimation de la petite séparation.

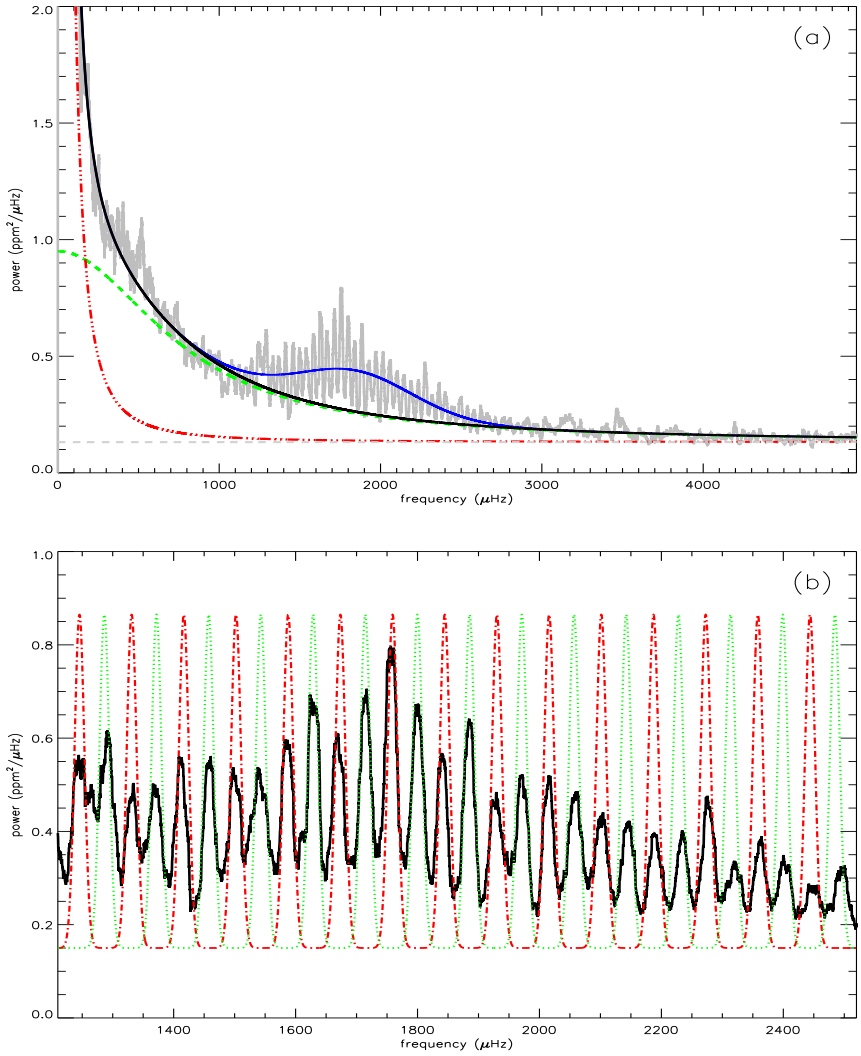


FIGURE 4.2 – (a) Exemple d'ajustement des profils de type Harvey superposés au spectre de puissance lissé de HD 49933. On distingue les composantes de Harvey associées à l'activité (pointillés rouges), à la granulation (pointillés verts) et le niveau de bruit blanc (pointillés gris). La somme des profils de bruit est en noir tandis que la ligne bleue représente le modèle dans son intégralité (enveloppe des modes + bruits). (b) Représentation des *a priori* sur les fréquences des modes pour le modèle M_A de l'étoile en question (Run IRa-01) ($l = 0$: pointillés rouge ; $l = 1$: pointillés vert). Ces courbes sont tiré de Benomar et al. (2009a).

TABLE 4.3 – Synthèse des scénarii envisagés dans un premier temps pour évaluer les capacités de l'algorithme TAMCMC développé et introduit au chapitre précédent. Ces simulations ne comportent que 3 modes ($l = 0$, $l = 1$, $l = 2$). Les scénarii s'inspirent fortement des problèmes rencontrés lors de l'ajustement par MLE de HD49933. L'exploration des deux identifications possibles s'effectue simplement en inversant les *a priori* sur les fréquences (ajustement d'un $l = 1$ sur le $l = 0 - 2$ et inversement). Pour ce qui concerne les sigles définissants la nature des *a priori* utilisés, se référer à l'Annexe A.

Paramètres	Valeurs vraies du cas, Favorable	Défavorable	*a priori* utilisés
$H_{l=0}$ (ppm^2/μHz)	1	1	$U[0,5]$
$H_{l=1}$ (ppm^2/μHz)	1.5	1.5	$U[0,5]$
$H_{l=2}$ (ppm^2/μHz)	0.5	0.5	$U[0,H_{l=0}]$
$\nu_{l=0}$ (μHz)	40	40	$U[30,50]$
$\nu_{l=1}$ (μHz)	70	70	$U[51,80]$
$\delta\nu_{02}$ (μHz)	$\delta^* = 8$	$\delta^* = 6$	$N(\delta*, 0.5\delta*)$
Γ (μHz)	2	5	$U[0,20]$
ν_s (μHz)	3.5	3.5	$N(3.5, 0.35)$
α (°)	30	30	$U[0,90]$
B_0 (ppm^2/μHz)	0.05	0.2	$U[0,0.5]$

Quatres scenarii ont donc été envisagés lors des premiers tests de l'algorithme. Nous avons considéré deux situations de rapport signal à bruit : un rapport H/N moyen de 5 et un autre de 20. Pour chacune de ces situations, nous nous sommes placés dans une situation dite "favorable" et une autre dite "défavorable". Dans le cas favorable, les grandeurs largeur (Γ), splitting (ν_s) et petite séparation sont choisis de telle sorte que les modes sont distants et résolus : $\delta\nu_{02} > \nu_s > \Gamma$. A l'inverse, dans le cas défavorable, on envisage un comportement des modes stellaires rendant difficile leur extraction. La largeur des modes a une valeur voisine de celle de la petite séparation et le splitting est inférieur à ces deux dernières. Ainsi, les modes ne sont que très difficilement discernables visuellement, même dans des conditions de rapport signal à bruit excellentes. L'angle d'inclinaison stellaire a été considéré identique pour toutes les simulations ($\alpha = 30°$). La Table 4.3 résume les quatre situations envisagées et les *a priori* associés. La résolution spectrale atteinte par CoRoT pour HD49933 étant de 0.19 μHz à l'époque où nous avons menés ces simulations (60 jours d'observations pour l'IRa-01), les simulations présupposent une résolution de 0.2 μHz (équivalent à une observation de 57.9 jours).

Au total, toutes situations confondues, 8 modèles ont été testés, 4 correspondants à une identification correcte des modes et les 4 autres à une identification incorrecte. Les termes de log-probabilité tels que définis par l'Eq.3.54 ont été calculés pour chacun d'eux avec 10 chaines parallèles et 200 000 itérations. On retrouvera ces valeurs et les probabilités relatives associées dans la Table 4.4 (calculées *via* l'Eq.2.10).

Pour simuler les propriétés du bruit, une manière simple consiste à définir un modèle de spectre de puissance $M(\nu)$ en se servant des paramètres cités Table 4.3 ; puis à calculer le module du spectre de puissance bruité $S(\nu)$ en se servant de la relation,

$$S(\nu) = \left| f_1(\nu)^2 \frac{M(\nu)}{2} + f_2(\nu)^2 \frac{M(\nu)}{2} \right|, \tag{4.5}$$

TABLE 4.4 – Table résumant les différentes log-probabilités et probabilités obtenues lors de simulations préliminaires mettant en oeuvre un ajustement sur 3 modes de bas degré ($l = 0, l = 1$, $l = 2$). Dans chacun des cas le modèle favorisé est le bon. Les variations typiques sont indiqués entre parenthèses : elles fournissent une indication sur l'erreur effectuée sur la probabilité. Une plage de probabilité est calculé en accord avec ces variations typiques de l'estimation.

H/N	Condition	log-probabilité (Mauvaise identification)	log-probabilité (Bonne identification)	probabilité
25	Favorable	671.3(±0.3)	681.49(±0.07)	$> 99.99\%$
25	Défavorable	390.14(±0.02)	392.06(±0.05)	$[86.6, 87.8]\%$
5	Favorable	167.45(±0.07)	168.90(±0.01)	$[79.9, 82.2]\%$
5	Défavorable	−52.10(±0.03)	−50.09(±0.3)	$[84.7, 91.1]$

L'excitation stochastique est ainsi simulée par deux réalisations f_1 et f_2 indépendantes de bruits gaussiens mimant le bruit agissant sur la partie réelle et imaginaire d'un signal.

Comme illustré par les spectres de la table de figures 4.5, dans des circonstances défavorables de largeur/splitting/petite séparation, c'est principalement l'asymétrie sur l'excès d'energie du couple $l = 0 - 2$ engendrée par la présence du $l = 2$ qui permet de distinguer ces deux modes. Ceci est valable, même à très haut rapport signal à bruit : une observation visuelle, ne permet que difficilement de définir lequel des excès de puissance peut être associé à un $l = 1$ et lequel correspond au couple $l = 0 - 2$. Seul un critère statistique permet d'établir de la manière la plus objective possible (ou tout du moins dans le cas d'une approche bayésienne, de manière reproductible) laquelle des identifications est la plus probable.

Sur ces mêmes figures, nous avons représenté les ajustements de spectre obtenues dans chacun des cas envisagés en nous basant sur deux critères distincts : la médiane (en bleu sur la figure) et la moyenne (en orange). Les différences ne sont pas significatives mais on remarque que la médiane fournit systématiquement des valeurs légèrement inférieures du paramètre. En fait, un tel critère est moins sensible aux aléas (ou erreurs) importantes. Par exemple, dans le cas d'une distribution *a posteriori* bi-modale, la médiane fournit une valeur plus robuste et généralement plus vraisemblable que la moyenne. Un bon exemple est illustré par la fréquence du mode $l = 2$ dans la table de figures 4.6. Elle est constituée de deux maxima dont un correspond en fait au mode rotationnel $m = +1$.

Par définition, la probabilité de trouver la bonne valeur du paramètre de part et d'autre de la position médiane est identique (50% de chance de chaque coté de la médiane). On peut aussi se demander si le maximum de probabilité est un critère fiable. En fait, il est difficile de l'estimer correctement car nous n'avons pas affaire à des fonctions analytiques de la fpd. Qui plus est, si nous nous trouvons dans une situation où la fdp est par exemple constituée de deux maxima, de probabilité voisine, lequel des deux prendre ? Tous ces arguments militent en faveur du choix de la médiane comme critère statistique pour définir la "valeur la plus probable" autour de laquelle les intervalles de confiance à 1σ et 2σ seront construits. Ceux ci sont obtenus par simple sommation à droite et à gauche de la médiane, tel que l'aire de part et d'autre de celle ci vaut respectivement $\frac{68.3}{2}\%$ et $\frac{95.5}{2}\%$.

Les fdp présentées dans les tables 4.6 et 4.7 sont très riches en informations et méritent que l'on s'y attarde. Nous ne présenterons que les deux cas extrêmes déjà discutés précédemment : le cas haut H/N dans des conditions stellaires favorables et le cas bas rapport H/N dans des conditions stellaires défavorables. La colonne de gauche est associée aux fdp obtenues lors de

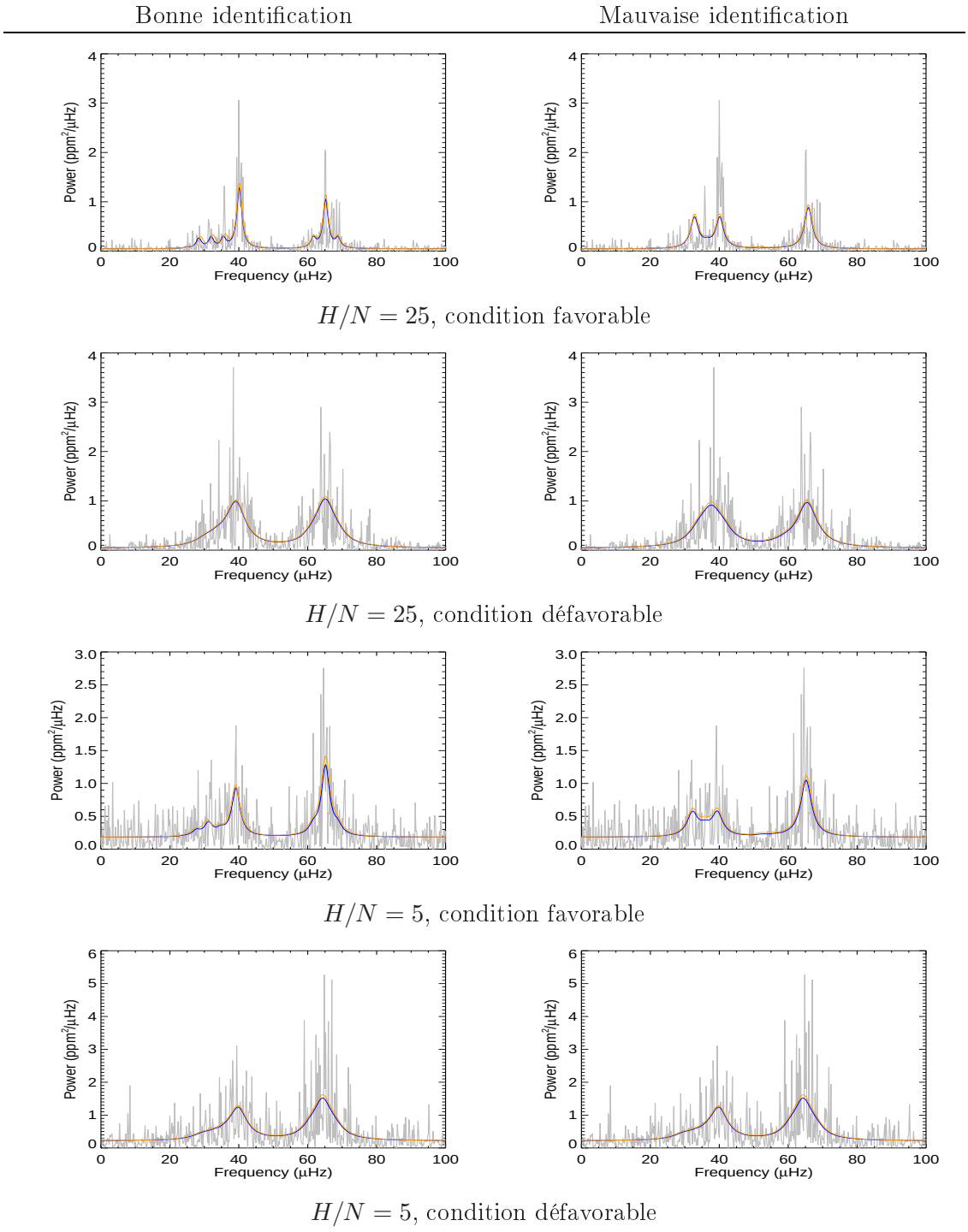
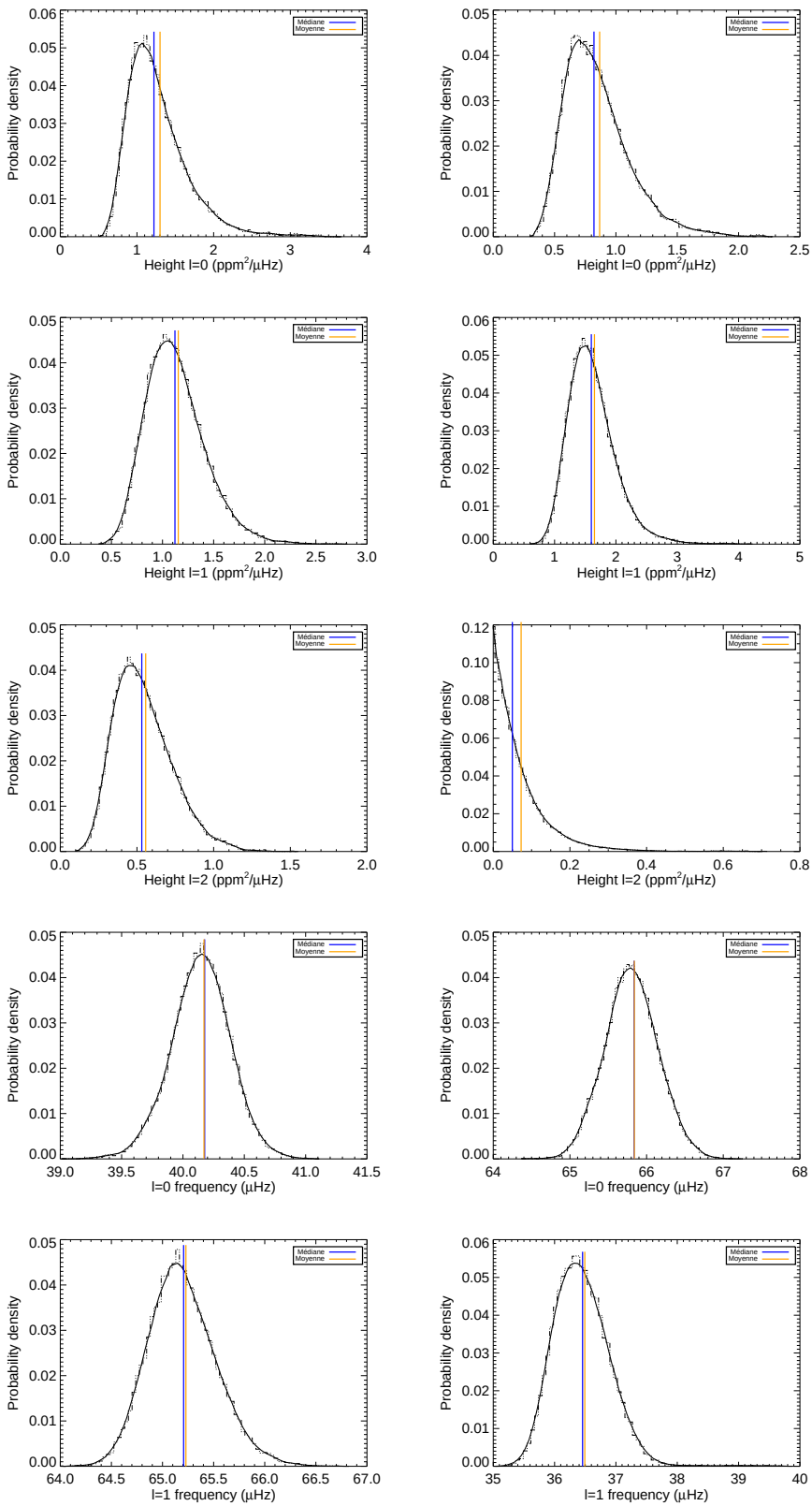


TABLE 4.5 – Spectres simulés contenant 3 modes de degré $l = 0$, $l = 1$, $l = 2$ et les ajustements correspondants : en bleu, l'ajustement selon la médiane des distributions *a posteriori*. En orange, selon la moyenne. La plupart du temps, les différences ne sont pas radicalement différentes.

Bonne identification | Mauvaise identification



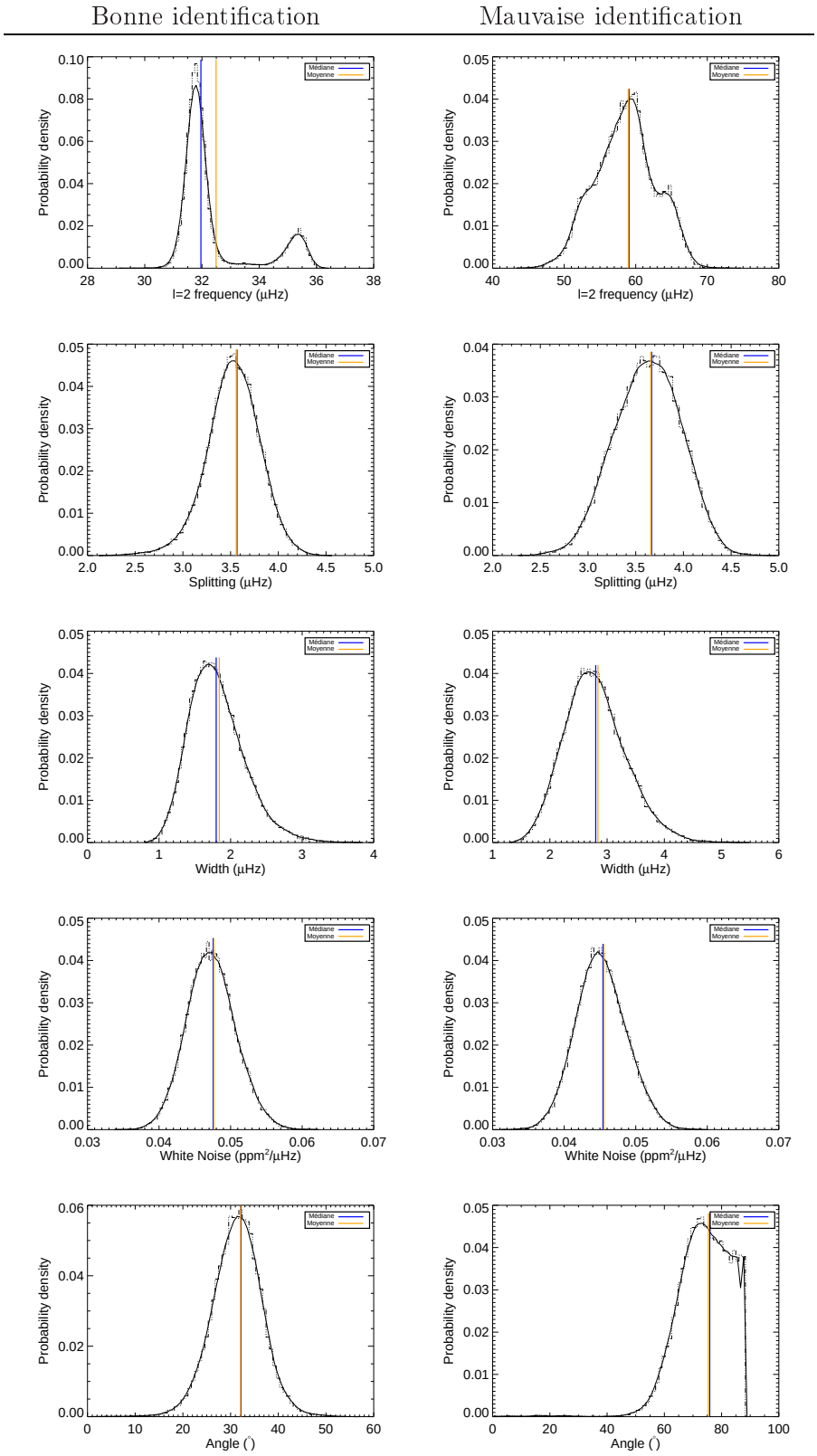
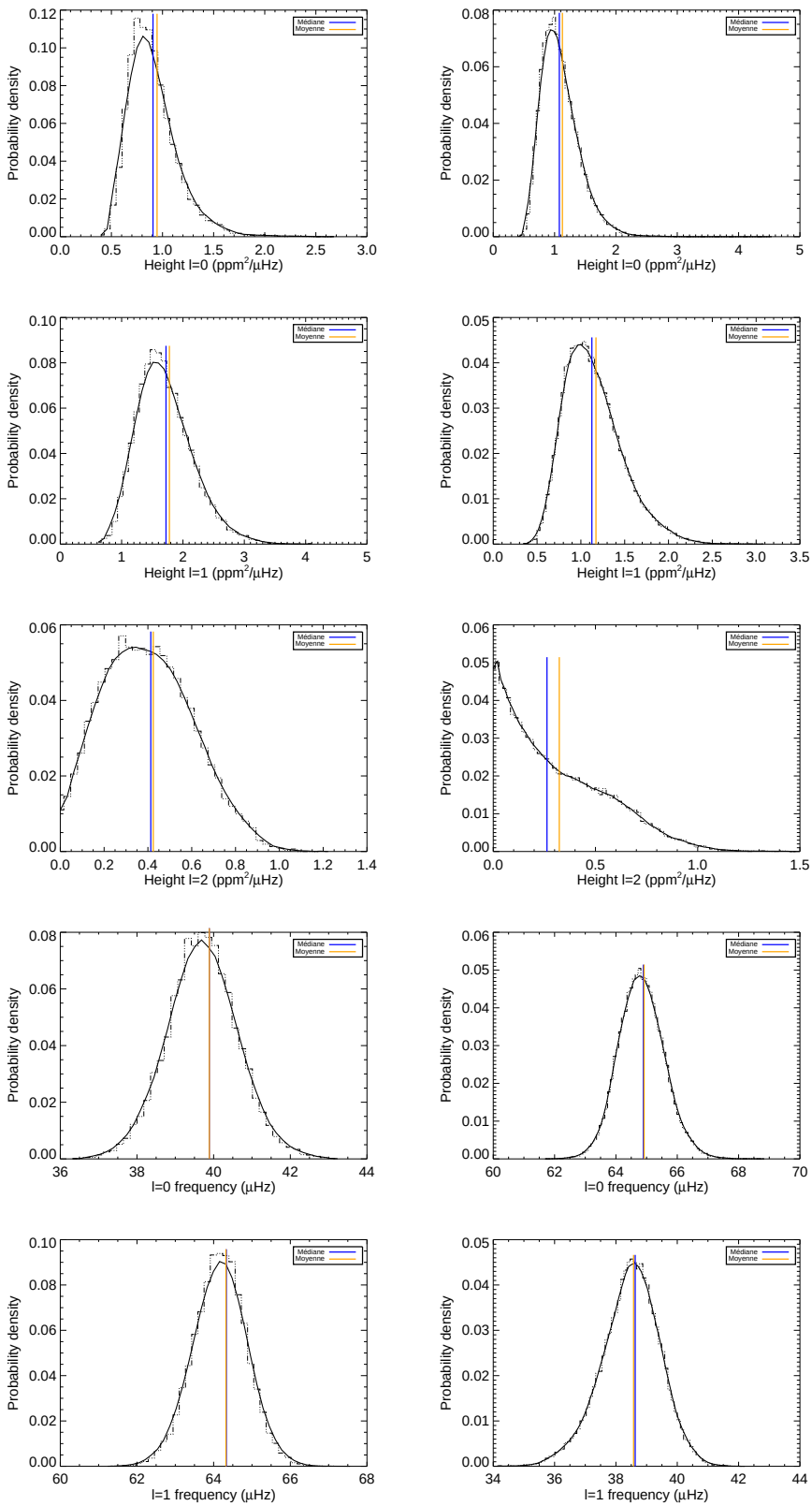


TABLE 4.6 – Fonctions de densité de probabilité pour les spectres simulés à un rapport $H/N = 25$, en condition favorable et pour les deux identifications. Ici la fréquence du $l = 2$ est bi-modale : le pic de faible probabilité correspond à un l'une composantes rotationnelle du mode. Les positions de la moyenne et de la médiane sont indiquées respectivement par des lignes oranges et bleues.

Bonne identification | Mauvaise identification



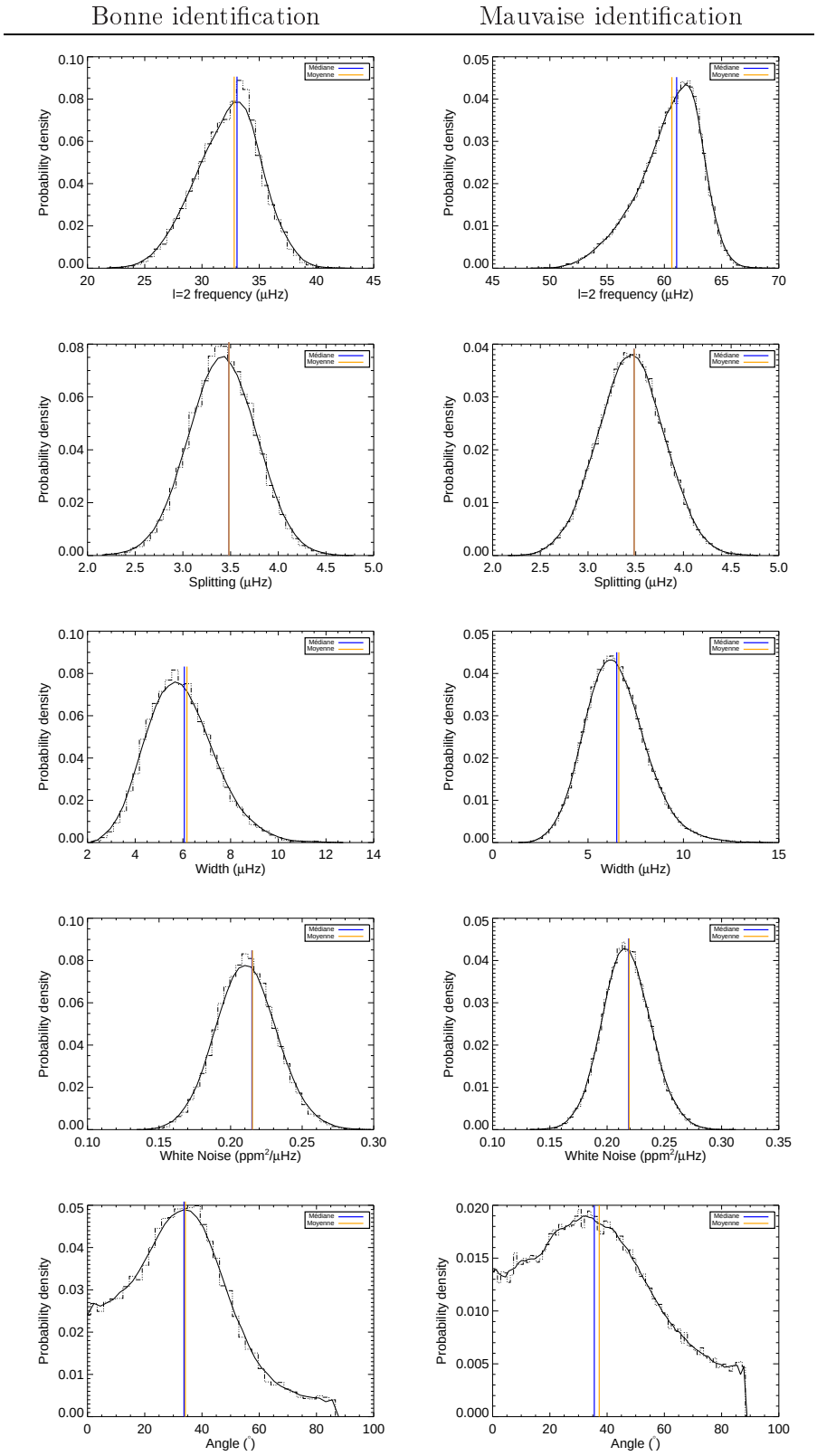


TABLE 4.7 – Fonctions de densité de probabilité associées aux spectres simulés pour un rapport $H/N = 5$ en condition favorable et pour chacun des ajustements possibles. Les positions de la moyenne et de la médiane sont indiquées respectivement par des lignes oranges et bleues.

l'ajustement du spectre par la bonne identification, la colonne de droite lors de l'ajustement par la mauvaise. La première chose que l'on remarque est la diversité des fdp obtenues : les hauteurs, largeurs et l'angle d'inclinaison ne suivent généralement pas une loi gaussienne tandis que les fréquences le sont plus volontiers. En fait, pour ce qui concerne les hauteurs et largeurs, ce fait est bien connu. Appourchaux et al. (1998), via des simulations Monte-Carlo a mis en évidence que la statistique de ces paramètres est relativement bien décrite par une fonction log-normale. Par contre, la nature parfois très dispersée (cf. cas à $H/N = 5$) de la fdp de l'angle d'inclinaison est une nouveauté. Un autre point important à relever concerne la nature de la fdp de la hauteur du mode $l = 2$. Lorsque l'on ajuste le spectre par la mauvaise identification (*i.e.* un couple $l = 0 - 2$ sur un $l = 1$ et vice-versa), la fdp de la hauteur du mode a la même allure que la statistique du bruit en χ^2_2. Indépendamment de la probabilité fournie par l'Eq.2.10, la nature de la fdp du $l = 2$ nous indique donc quelle est en théorie la bonne identification. Cependant, il faut tout de même nuancer ce fait car la superposition des modes de degré pair peut modifier cette distribution. Ce phénomène est illustré par les figures de la Table 4.7, dans le cas de conditions stellaires défavorables (visible même à haut rapport H/N). Dans cette situation, la médiane est un bien piètre indicateur car sa valeur peut être similaire que l'ajustement soit le bon ou pas. Enfin, on fera remarquer que la superposition du $l = 2$ sur le $l = 0$ induit une sous-estimation de la hauteur relative du $l = 2$. Ce fait sera tout aussi net lorsque nous présenterons les résultats de simulation sur un nombre de modes important et prenant en compte les propriétés du bruit stellaire (sous-chapitre suivant).

Pour ce qui concerne le splitting, la contrainte imposée (10% de la valeur attendue) semble relativement forte car la fdp sur ce paramètre est dominé par l'*a priori* : la valeur moyenne et l'écart type sont très proches des valeurs tolérées et ce, même à très bon rapport H/N. Il semblerait donc que le spectre ne contienne pas assez d'information pour compenser l'effet de l'*a priori*.

Dans la Table 4.4 on a résumé les différentes probabilités obtenues dans chacun des scénarii envisagés. On remarquera que systématiquement nous obtenons la bonne identification. Les conditions défavorables ont un impact significatif sur la capacité de discrimination des identifications à haut rapport signal sur bruit. Mais les conditions défavorables ont moins d'impact à bas H/N. La probabilité semble être même plus élevé dans le cas défavorable que dans le cas favorable. Toutefois il faut relativiser cette affirmation car ce test n'a été effectué qu'en se basant sur une seule réalisation bruité du spectre. Cette réalisation peut être exceptionnellement en faveur du bon modèle.

L'échantillonnage par MCMC se révèle être un outil très efficace de diagnostic du signal stellaire. Nous avons montré qu'il était probablement possible de discriminer entre deux ajustements du spectre de puissance d'une étoile, à la fois par l'observation de l'allure de la fdp de la hauteur du $l = 2$ et par le calcul de probabilité relative. Néanmoins dans le cas d'un signal stellaire d'une étoile de type solaire dans la séquence principale, l'information spectrale est plus riche en terme de modes et de bruit, mais aussi plus complexe à analyser. Pour confirmer et évaluer les tendances que nous avons présentées dans ce chapitre, nous avons aussi mené des analyses sur simulations reproduisant dans sa globalité un spectre d'étoile. En effet, il est avantageux d'ajuster les spectres stellaires dans leur intégralité en une seule passe car on peut imposer aux modes individuels des dépendances à des paramètres stellaires globaux. Par exemple l'angle d'inclinaison stellaire et sous certaines hypothèses, le splitting rotationnel constant. Ou encore le profil de bruit qui, traité d'un seul tenant permet une bien meilleur estimation de celui ci. Tous ces paramètres globaux affectant tous les modes, un ajustement global améliore naturellement la qualité de l'ajustement (à condition que la complexité du modèle ne croisse pas plus vite que l'accroissement d'information présente dans le spectre).

4.2.2 Spectres stellaire simulés

Dans cette section, nous allons présenter les résultats principaux d'une étude plus réaliste du comportement des modes stellaires. Nous nous limiterons à l'étude des propriétés des modes p dont la répartition approximative dans le spectre de puissance est fournie par la loi asymptotique.

La richesse d'information présente dans un spectre de puissance pris dans sa globalité nous ont poussé à définir un protocole d'analyse en plusieurs étapes, que nous décrirons. Les simulations menées ont pour but d'évaluer :

- la pertinence du protocole d'analyse choisi dans des situations variées.
- l'influence que peut avoir un mauvais choix de l'*a priori* sur la fdp et sur la probabilité des modèles.

Soulignons tout de même que ces simulations ne constituent en aucun cas une preuve de fiabilité de la méthode pour de véritables données. En effet, globalement on peut définir 4 situations :

1/ Les hypothèses physiques effectuées lors des simulations sont valides et les hypothèses sur les propriétés statistiques du signal sont valides.

2/ Les hypothèses physiques effectuées lors des simulations sont valides mais les hypothèses sur les propriétés statistiques du signal ne sont pas valides. Ceci peut être le cas si des termes de bruit (par exemple instrumentaux) ont été négligés.

3/ Les hypothèses physiques effectuées ne sont pas valides mais les hypothèses sur les propriétés statistiques du bruit le sont.

4/ Enfin, les hypothèses tant sur la physique du phénomène que sur les propriétés statistiques ne sont pas valides.

Au regard de cette liste, on comprend bien vite que les simulations ne permettent que d'évaluer la situation 1. En règle générale, les aléas de diverses natures sont bien connus et la situation 2 n'est pas problématique car une fois qu'ils sont identifiés, nous pouvons trouver un moyen de nous en affranchir au mieux. Soulignons par exemple que dans le cas de CoRoT, la statistique suivit par le bruit peut être influencée par la présence de pics orbitaux. En effet, le satellite est en orbite polaire basse terrestre. Or il existe une zone appelé *anomalie sud-atlantique* (SAA) pour laquelle la ceinture de Van Allen est la plus proche de la surface terrestre. Le satellite étant à une altitude invariante (896 km), à chacun de ces passages dans cette zone il subit un flux de radiations anormalement élevé. Ce phénomène se produit selon une modulation de période égale à la période orbitale. La durée du passage dans la zone est relativement variable (fonction de l'extension spatiale de l'anomalie au moment du passage) et influe sur l'intensité du signal parasite. Les CCD subissent l'impact des particules piégées dans la SAA ce qui peut fortement perturber les courbes de lumières. Au final, malgré les diverses traitements que subissent les données en vue de limiter les effets instrumentaux de la SAA, un pic parasite à 161.57 μHz et ses harmoniques se manifestent dans le spectre de puissance. La fréquence du fondamental correspond exactement à la période orbitale du satellite de 103 min.

Un autre signal parasite remarquable est dû à la dilatation et contraction par effet thermique de la structure du satellite lors de son passage de la zone diurne à la zone nocturne et inversement. Les gradients de température peuvent être importants dans l'espace lors du passage de l'ombre à la lumière. Qui plus est, ces transitions sont associées au basculement de l'alimentation des instruments du solaire vers une batterie (ou inversement). L'alimentation électrique n'est alors pas parfaitement stable dans cette phase. Enfin, ces perturbations sont parfois modulées par la rotation de la terre. Ces phénomènes se manifestent par deux raies plus ou moins intenses selon le domaine spectral, encadrant a ± 11.57 μHz (et ses harmoniques) le pic dû à la SAA.

La seule solution viable consiste à ne pas tenir compte des points du spectres impactés par ces effets lors des ajustements. Ainsi en théorie, la statistique du spectre de puissance en χ^2_2 est

préservée. Les problème liés à la SAA sont inexistants dans la mission Kepler car le satellite n'est pas en orbite terrestre mais suit une trajectoire approximativement parallèle à celle-ci autour du soleil. Cependant, il subit un flux de radiation moyen supérieur à CoRoT. La plupart des effets d'impacts sont corrigibles et corrigés.

Pour ce qui concerne les situations 3 et 4, nous ne pouvons que nous reposer sur la méthode scientifique pour invalider/valider les hypothèses physiques énoncées. Tout le travail scientifique est d'identifier ces situations à l'épreuve des faits et ainsi, réviser de manière progressive nos connaissances.

Conditions et résultats des simulations

Pour tester les principes généraux de la méthode présentée Chapitre 4.1.2 nous avons mené des simulations plus complètes que les précédentes. Au total c'est 6 scénarii que nous avons envisagés. Ils diffèrent encore une fois de par les niveaux de bruit et la nature du signal stellaire. La plupart des scénarii (1 à 5) évalue la possibilité d'identifier correctement les 3 degrés les plus bas des modes. L'évaluation porte sur l'ajustement global de 8 ordres radiaux.

Le scénario 6 considère le cas d'un spectre de puissance avec seulement les deux plus bas degrés, $l = 0$ et $l = 1$. Il est conçu pour tester les capacités de discrimination de modèles par calcul de probabilité en l'absence de $l = 2$ dans le spectre synthétique. Lors de la comparaisons de modèles, 4 possibilités sont envisagées et testées. On note M_k^l les modèles associés à ces possibilités. L'indice supérieur désigne le degré l maximum ajusté. L'indice inférieur correspond à l'identification. Par exemple, le modèle M_A^1 correspond à l'identification A pour laquelle on ne suppose que l'existence des modes $l = 0$ et $l = 1$ dans le spectre de puissance. C'est donc la bonne identification avec le bon nombre de degré.

Tous les spectres synthétiques supposent que la hauteur relative est égale aux valeurs solaires : $V_{l=1} = 1.5$, $V_{l=2} = 0.5$. Le profil de bruit est constitué d'une semi-lorentzienne et d'un bruit blanc. La résolution spectrale est définie à 0.2 μHz tout comme dans les simulations présentées Chap. 4.2.1. Les fréquences sont distribuées selon la loi asymptotique avec $\Delta\nu = 90$ μHz et l'angle d'inclinaison est fixé à 40°.

Pour chacun des scenarii, on a fait varier la hauteur, la largeur, le splitting rotationnel et la petite séparation $\delta\nu_{02}$. Au final, on peut distinguer 4 classes de scenario :

Situation 1. Une hauteur des modes relativement haute au regard du bruit (H/N $>$ 30) avec un $l = 2$ clairement distinct du $l = 0$. Le splitting rotationnel est très faible comparé la petite séparation ($\nu_s << \delta\nu_{02}$). Ces conditions simulent une étoile dans laquelle les multiplets des modes sont distincts (Scénario 1).

Situation 2. Similaire au Scénario 1, mais à plus haut niveau de bruit. $H/N \simeq 4$ (Scénario 2).

Situation 3. Un rapport H/N intermédiaire (3 à 10) entre les scénarii 1 et 2. On a $\delta\nu_{02} \simeq \nu_s$ ce qui provoque une superposition des composantes rotationnelle du $l = 2$ sur le $l = 0$. Les largeurs des modes sont accrues comparativement aux scénarii 1 et 2, rendant plus difficile encore la séparation des modes voisins (Scénarii 3 à 5).

Situation 4. Le scénario 6 nous sert à déterminer les possibilités de comparaison offertes par l'approche bayésienne entre modèles de taille d'espace des paramètres différents.

On ajoutera que les *a priori* sur le splitting et sur la petite séparation des scénarii 1 et 2 suppose que nous connaissons la valeur vraie de ces paramètres. Dans le cas des scénarii 3 et 5 on utilise un *a priori* décentré de la valeur vraie du paramètre pour évaluer l'influence d'un mauvais choix d'*a priori*. Les différentes valeurs utilisées lors des simulations sont résumées Table 4.8 (valeurs utilisées pour générer les spectres de puissance simulés) et Table 4.9 (les *a priori*).

TABLE 4.8 – Paramètres ayant servit à la génération des spectres simulés : Deux profils de variations de la hauteur en fonction de la fréquence ont été envisagés : une hauteur constante pour tous les modes (C) ou un profil gaussien, plus réaliste (G).

Scénario	$\delta\nu_{02}$ (μHz)	ν_s (μHz)	Γ (μHz)	H/N ($l = 0$)	H Profil
1	8	2	[1,5]	> 30	C
2	8	2	[1,5]	$\simeq 4$	C
3	5	3.5	[2,10]	$\simeq 6$	C
4	5	3.5	[3,10]	3-10	G
5	5	3.5	[4,13]	3-8	G
6	/	3.5	[3,10]	3-8	G

TABLE 4.9 – *A priori* utilisés lors de l'analyse des spectres simulés. L'*a priori* sur l'angle d'inclinaison i est uniforme. Les autres paramètres sont contraints par une gaussienne.

Scénario	i	ν_s	$\Delta\nu$	$< \delta\nu_{02} >_{th}$
1	$U[0-90]$	$N(2, 0.2)$	$N(90.2, 0.5)$	N$(8, 4)$
2	$U[0-90]$	$N(2, 0.2)$	$N(89.9, 0.6)$	$N(8, 4)$
3	$U[0-90]$	$N(3.5, 0.35)$	N$(90.2, 0.6)$	$N(8, 4)$
4	$U[0-90]$	$N(3.5, 0.35)$	$N(90.0, 0.3)$	$N(5, 2.5)$
5	$U[0-90]$	$N(3.5, 0.35)$	$N(90.0, 0.4)$	$N(6, 3)$
6	$U[0-90]$	$N(3.5, 0.35)$	$N(90.1, 0.5)$	$N(6, 3)$

En les comparant, le lecteur pourra juger plus quantitativement les situations envisagées.

Pour chacun des scénarii on acquiert 1 000 000 d'échantillons une fois la phase de chauffage passée. La Table 4.10 est une synthèse des rapport concurrentiels obtenus. Dans tous les cas, le coefficient concurrentiel est en faveur de la bonne identification. Pour les scénarii 1 à 3, l'identification est correcte à plus de 99.99%. On remarquera que la nature du profil des modes, la superposition engendrée par la proximité entre le $l = 0$ et le $l = 2$ et une largeur importante des modes réduit fortement le rapport concurentiel. Dans les conditions les plus dégradées (scénario 5), le rapport concurentiel devient même ambigu et on est plus capable de décider objectivement laquelle des deux identifications est la bonne (rapport concurentiel inférieur à 3). Dans le cas du scénario 6, les *a priori* sur la petite séparation $\delta\nu_{02}$ agissent selon nos attentes en pénalisant le modèle le plus complexe. Mais nous notons que le bon modèle (M_A^1) n'est affecté que d'une probabilité de 64% et dans une situation réelle aucune conclusion définitive n'est possible.

La distribution de la hauteur relative du $l = 2$ peut varier fortement d'une identification à l'autre. La fdp de ce paramètre suit la distribution du bruit dans la mauvaise identification mais est plutôt gaussienne dans le cas de la bonne. On notera toutefois que celle ci est sous-estimée d'environ 30% lorsque les modes se superposent. Tout ceci confirme les résultats des simulations préliminaires décrites Chap.4.2.1 pour lesquels nous n'ajustions que 3 modes. On notera aussi que lorsque l'*a priori* sur la petite séparation est incorrect, elle est relativement mal estimée. Par exemple pour le scénario 5, la valeur médiane obtenue est 6.1 ± 0.7 μHz ce qui est environ 1 μHz au dessus de la valeur vraie. Ce décalage reste inférieur à 2σ. On remarquera que bien que les fréquences soient relativement bien contraintes (l'erreur typique tend vers la résolution

TABLE 4.10 – *Haut :* Rapports concurrentiels O_{M_A,M_B} pour les scénarii 1 à 5. *Bas :* rapport concurrentiel pour le scénario 6, consistant en une comparaison de 4 modèles de spectre ; deux avec $l = 2$, deux autres sans. Les valeurs de la matrice correspondent aux combinaisons de rapports concurrentiels possibles. On remarquera que le bon modèle est trouvé, mais seulement avec une probabilité de 64%. En situation réelle, nous ne pourrions que difficilement conclure.

Scenario	O_{M_A,M_B}
1	2×10^{48}
2	7×10^{26}
3	1.5×10^{4}
4	7.209
5	1.578

Scenario 6					
$M_k^l \backslash M_{k'}^{l'}$	M_A^1	M_B^1	M_A^2	M_B^2	$P_R(M_k^l \vert y, I)$
M_A^1	1	4.6104	92.6707	2.9385	64%
M_B^1	0.2169	1	20.1003	0.6374	13%
M_A^2	0.0317	0.0495	1	0.0108	1%
M_B^2	0.3403	1.5690	31.5373	1	22%

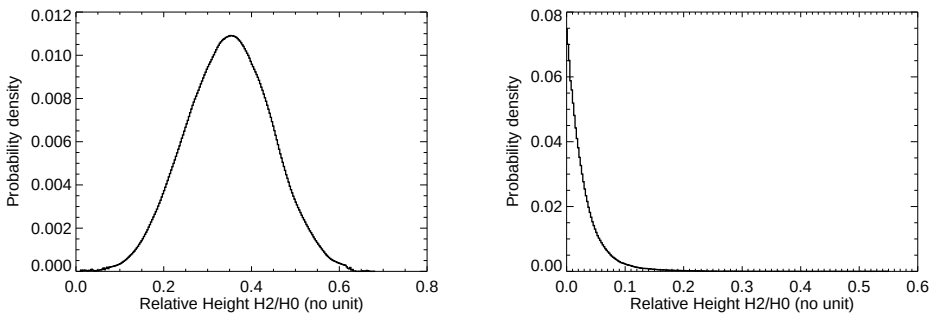


FIGURE 4.3 – Exemple de hauteur relative des modes $l = 2$ dans le cas des simulations (scénario 3) pour la bonne identification (a) et la mauvaise (b). Dans le cas de la bonne identification la hauteur relative $V_{l=2}$ est sous-estimé de 30%. Dans le cas de la mauvaise identification, la hauteur relative du mode semble suivre la distribution statistique du bruit. Ceci confirme les simulations menées sur seulement 3 modes.

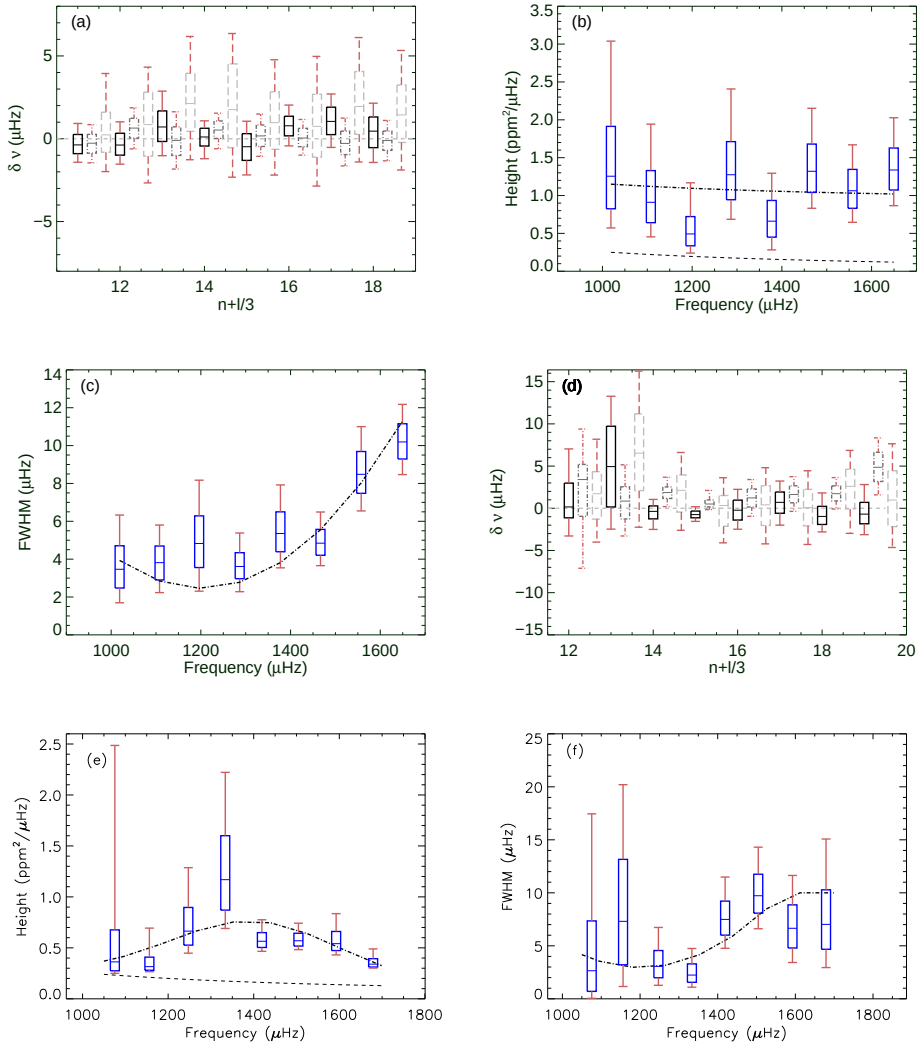


FIGURE 4.4 – Comparaison entre les valeurs vraies et les valeurs obtenues par ajustement pour les spectres simulés. Deux scénarii sont présentés : scénario 3 (a,b,c) et scénario 4 (d,e,f). On utilise ici une représentation en "boite à moustache" (*whisker plot*). Le rectangle central correspond à un intervalle de confiance à 68.3% (équivalent à 1 σ) autour du médian (trait à l'interieur du rectangle). Les lignes de part et d'autre du rectangle indiquent l'intervalle de confiance à 95.5% (équivalent à 2 σ). Les boites à moustaches sont habillées de couleurs pour les frequences (a,d). La ligne pleine noire représente le $l = 0$. La ligne grise en pointillés est associée aux $l = 1$. Le $l = 2$ est dans un gris moins foncé. L'axe des abcsisses est défini par la quantité $n + l/3$. Le biais statistique dû à un choix inexact d'*a priori* est visible (scénario 3) sur les fréquences du $l = 2$. Dans le cas des hauteurs (b,e) et largeurs (c,f), on compare les valeurs vraies utilisées lors de la génération du spectre (lignes pointillés) aux valeurs estimées. Le profil de bruit est aussi représentée.

spectrale dans les cas les plus favorables) ce n'est pas du tout le cas des hauteurs et largeurs. L'anticorrélation entre hauteur et largeur est aussi assez bien visible sur ces figures. On signalera que tous les résultats présentés incluent les valeurs vraies à tout au plus 2σ.

La conclusion de ces simulations est qu'à condition d'ajuster des modèles de spectres dont les hypothèses physiques sont *exactes* (profils lorentziens symétriques, splitting constant, rotation lente, ...), il est très probablement possible d'évaluer de manière correcte quel est la bonne identification pour la plupart des étoiles et ce, même si les conditions de rapport hauteur des modes sur bruit ($H/N \gtrsim 3$) et/ou les propriétés stellaires sont mauvaises.

4.3 Applications aux étoiles

Maintenant que nous avons un degré de confiance suffisant en notre algorithme et en ses capacités, nous pouvons l'appliquer à de véritables cas stellaires.

4.3.1 Un cas stellaire intensivement étudié : HD49933

HD 49933 est une étoile qui a été observée une première fois par CoRoT tout juste après son lancement durant 60 jours (IRa-01). Les données se sont révélées assez difficiles à analyser comme nous pourrons le voir dans ce qui suit. C'est pourquoi il a été décidé qu'elle soit réobservée durant 130 jours environ (LRa-01). Au final, les deux observations d'une durée totale de plus de 190 jours ont permis d'atteindre une duré d'observation inégalée pour cette mission.

Avant CoRoT, HD49933 avait déjà été observée depuis le sol par Mosser et al. (2005). Les données du IRa-01 ont été intensivement étudiées. Celles ci ont été en premier lieu analysées par Appourchaux et al. (2008) en utilisant une méthode par MLE. Nous les avons ré-analysées une première fois de manière succincte par une approche bayésienne en 2008, publié dans un compte rendu de conférence Benomar (2008). Une autre étude de ces données sismiques a été effectuée par Gruberbauer et al. (2009) mettant en oeuvre une approche Bayésienne et un modèle physique fortement simplifié (notamment absence de splitting rotationel et largeurs des modes constantes).

HD49933 est une étoile de la séquence principale de type spectral F5V. Ses caractéristiques principales sont : une masse approximative de $1.2M_\odot$ (Mosser et al. 2005) pour un rayon de $R/R_\odot = 1.34 \pm 0.06$ (Thévenin et al. 2006). Différentes estimées de $v sin\, i$ sont disponibles, allant de 9.5 ± 0.3 $km.s^{-1}$ (Mosser et al. 2005) à 10.9 $km.s^{-1}$ (Solano et al. 2005). Comme souligné par Appourchaux et al. (2008), le spectre puissance à basse fréquence semble porter la signature de la période rotationnelle causée par le transit de structures de surfaces. Ceci se manifeste par un pic significatif à 3.4 μHz. En se basant sur le $v sin\, i$, le rayon et la position de ce pic, l'angle d'inclinaison de l'étoile semble être compris entre $22 - 30°$, alors qu'en exploitant les données astérosismiques, Appourchaux et al. (2008) trouve $50 - 62°$. Pour ces auteurs, cette incohérence viendrait de la superposition des modes de même parité. Mais comme démontré plus loin, cette incohérence est simplement due à l'echec du test de rapport de vraisemblance au sens du MLE (se référer au Chap.2.1 pour une description du MLE). On comprend intuitivement que les solutions possibles (maxima de probabilité locaux) sont abondantes dans un spectre décrit par près d'une centaine de paramètres. Avec le recul permis par les données supplémentaires du LRa-01, il est évident que les résultats publiés dans cette étude sont incorrects car l'identification trouvée n'est pas la bonne. Dans tout ce qui suit, l'identification trouvée par Appourchaux et al. (2008) sera dénommée A tandis que l'identification opposée sera dénommée B.

Dans ce qui suit, nous reproduisons les résultats déjà publiés dans Benomar et al. (2009a) portant sur le IRa-01 et dans Benomar et al. (2009b) portant sur l'étude conjointe de IRa-01 et

LRa-01. On montrera comment l'analyse bayésienne intensive que nous avons effectuée sur cette étoile a permit d'établir statistiquement qu'elle est la bonne identification.

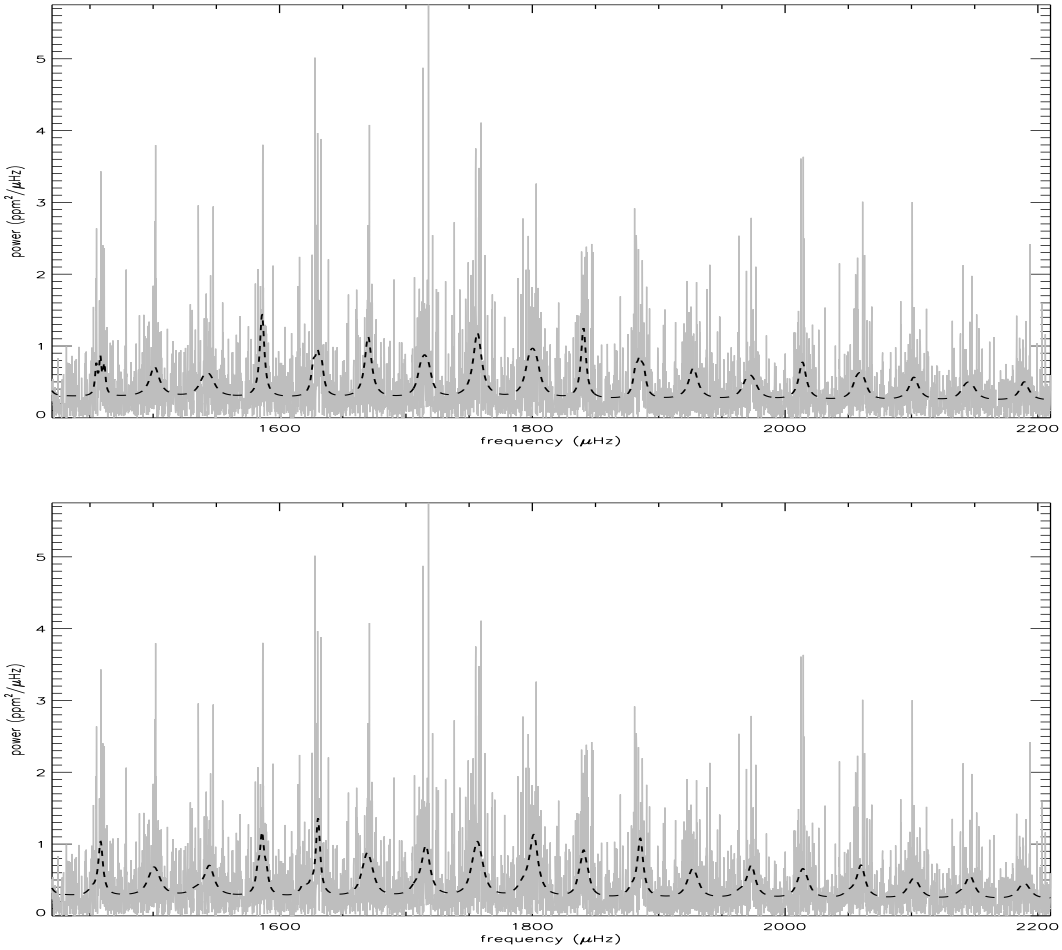


FIGURE 4.5 – *(Haut)* Spectre de puissance de HD49933 (IRa-01) superposé au modèle M_A^1 (modèle le plus probable). *(Bas)* Spectre de puissance de HD49933 (IRa-01) superposé au modèle M_B^2 (deuxième modèle le plus probable)

Sur la base des données du IRa-01 (cf. spectres Fig.4.5 ou diagramme échelle du spectre Fig.4.6) il est impossible de trancher quant à l'identification. Comme on l'a déjà dit, les modes de cette étoile sont bien plus larges que dans le soleil ce qui engendre une superposition à la fois des multiplets et du $l = 0$ sur le $l = 2$. Au final, toute observation visuelle des données sous quelques forme que ce soit ne permet pas d'aboutir à une conclusion quant à l'identification des excès d'énergie manifestes dans le spectre de puissance.

Comme pour toutes les étoiles que avons analysées, nos ajustements ont suivi la méthodologie décrite Chap.4.1. Pour HD 49933, nous avons comparé un total de 4 modèles comme pour le scénario 6 des simulations (2 modèles sans $l = 2$, deux supposant leur existence). Pour chacun d'eux, nous avons acquis 2 000 000 d'échantillons ce qui a garanti une précision de l'ordre de 0.2 dex sur les log-probabilités des modèles (soit environ 10-20% d'erreur estimée si les probabilités relatives à deux modèles sont égales). Les probabilités obtenues sont résumées Table 4.11 lorsque l'on considère les hauteurs relatives des modes libres. Le modèle le plus probable n'inclut pas

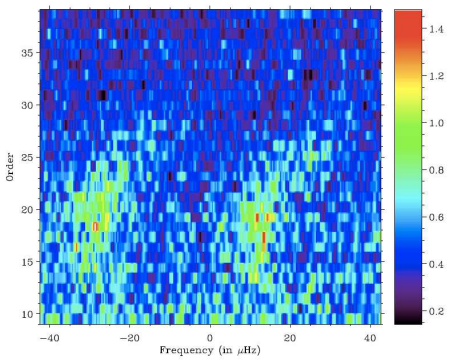


FIGURE 4.6 – Diagramme échelle du spectre de puissance pour le IRa-01. Ce diagramme échelle est tiré de Appourchaux et al. (2008)

TABLE 4.11 – rapport concurentiels et estimation des probabilités lors de l'analyse de HD49933 (IRa-01).

$M_k^l \backslash M_{k'}^{l'}$	M_A^1	M_B^1	M_A^2	M_B^2	$P_R(M_k^l \vert y, I)$
M_A^1	1	492.72	4.4817	2.7183	62.8%
M_B^1	0.0020	1	0.0091	0.0055	0.2%
M_A^2	0.2231	109.94	1	0.6065	14%
M_B^2	0.3679	181.26	1.6487	1	23%

de $l = 1$ et correspond à l'identification fournie par Appourchaux et al. (2008). La probabilité fournie ne favorise que très peu cette identification. Si nous calculons la probabilité marginale que l'identification A soit vraie, indépendament de la présence ou de l'absence du $l = 2$, on obtient,

$$P(M_A|D,I) = P_R(M_A^1|D,I) + P_R(M_A^2|D,I) = 63\%(\text{approx.} \pm 15\%) \quad (4.6)$$

$$P(M_B|D,I) = 1 - P_R(M_A|D,I) = 37\%. \quad (4.7)$$

Enfin, si on suppose comme totalement improbable le fait qu'il n'y ai pas de $l = 2$ dans le signal (ce qui est à l'opposé de ce que suggère le modèle le plus probable) alors on obtient,

$$P(M_A|D,I) = \frac{P_R(M_A^2|D,I)}{P_R(M_A^2|D,I) + P_R(M_B^2|D,I)} = 37.8\% \quad (4.8)$$

$$P(M_B|D,I) = 1 - P_R(M_A|D,I) = 62.2\% \quad (4.9)$$

Quelque soit le point de vue à partir duquel on aborde les données et contrairement à l'approche simplifiée par MLE, nous ne pouvons trancher quant à l'identification la plus probable : il n'y a pas de modèle qui se détache clairement des autres.

Notons que nous avons aussi essayé d'imposer une hauteur relative fixe au $l = 2$. Cela fournit des résultats très similaires à ceux présentés ici. En fait, les données du IRa-01 sont tout simplement trop pauvres d'un point de vue statistique pour pouvoir trancher. D'un point de vue décisionnel nous devons considérer que c'est M_A^1 qui est le plus probable. Mais d'un point de vue physique nous pensons que le $l = 2$ *doit* être présent et donc que c'est M_B^2 le modèle le plus probable, mais pas de manière significative. Au lecteur intéressé par plus de détails sur cette étude, nous conseillons de nous référer à Benomar et al. (2009a).

C'est ce résultat ambigu qui a motivé l'étude conjointe du IRa-01 et LRa-01. L'analyse conjointe des deux observations s'est effectuée après calcul d'un spectre de puissance moyen de résolution nominale de 60 jours. La durée de 60 jours de la première observation nous a obligée à un découpage en 3 morceaux (3×60 jours) et ainsi, nous avons dû négliger une dizaine de jours de données. Les tests que nous avons menés ont montré que les résultats ne variaient que très peu en fonction de la section de la série temporelle négligée. Les variations typiques sont systématiquement très inférieures aux barres d'erreurs à 1σ selon que l'on néglige les premiers jours ou les derniers jours d'observations.

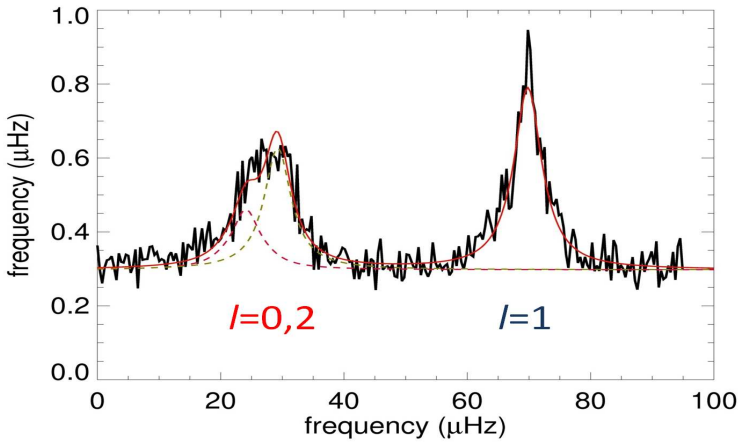


FIGURE 4.7 – Collapsogramme obtenue par collapse de 12 "tranches" de spectres.

Cette nouvelle analyse s'est effectuée sur le même principe que celle de Appourchaux et al. (2008). Plusieurs équipes ont analysé cette étoile de manière indépendante (la plupart n'ont analysé que le LRa-01 seul). Les conclusions de chacun ont été synthétisés dans Benomar et al. (2009b) et sont univoque. En ce qui concerne l'identification des modes, la caractéristique principale de cette étude et que presque tous les ajusteurs [3] utilisant des méthodes variées (MLE/MAP/MCMC) aboutissent à la même conclusion : la bonne identification est la B (opposée à celle précédemment trouvée). De plus, l'étude par MCMC que nous avons mené a évalué l'éventualité d'une absence de $l = 2$ (modèles M_x^1). Nous avons aussi considéré différents scénarii concernant la hauteur relative des modes. Au final, toutes les situations envisagées aboutissent à la même conclusion : la présence du $l = 2$ est avérée. Ceci correspond donc au modèle M_B^2. Les Fig.4.8, 4.9, 4.10 résument les différentes éventualités testées en ce qui concerne le bruit et les *a priori* sous forme d'un diagramme de décision. Ce type de diagramme permet de représenter graphiquement une chaine de processus de décision. Même si il existe un cas dans lequel la probabilité associée aux modèle dominant est relativement faible (la plus basse valeur obtenue est 81%), on insistera sur le fait que le résultat produit pointe toujours dans la même direction. L'identification étant établit avec une quasi-certitude, nous ne la discuterons plus.

Face à ces résultats, nous avons aussi cherché à évaluer la présence du $l = 3$. En théorie, nous pouvons utiliser la comparaison de modèles pour cela, mais la durée de calcul est très supérieure lorsque ce dernier est ajouté (de l'ordre de 50 % plus longue). Les temps de calcul par modèle étant déja de l'ordre de 3 semaines (3 à 4 millions d'échantillons), nous avons cherché à faire plus simple. Sur la base du collapsograme (cf. Fig.4.7) nous avons cherché à établir par ajustement si un modèle tenant compte du $l = 3$ était probant (hauteur relative du $l = 3$ fixé à la valeur

3. Une personne ne pouvait conclure : Rafael Garcia

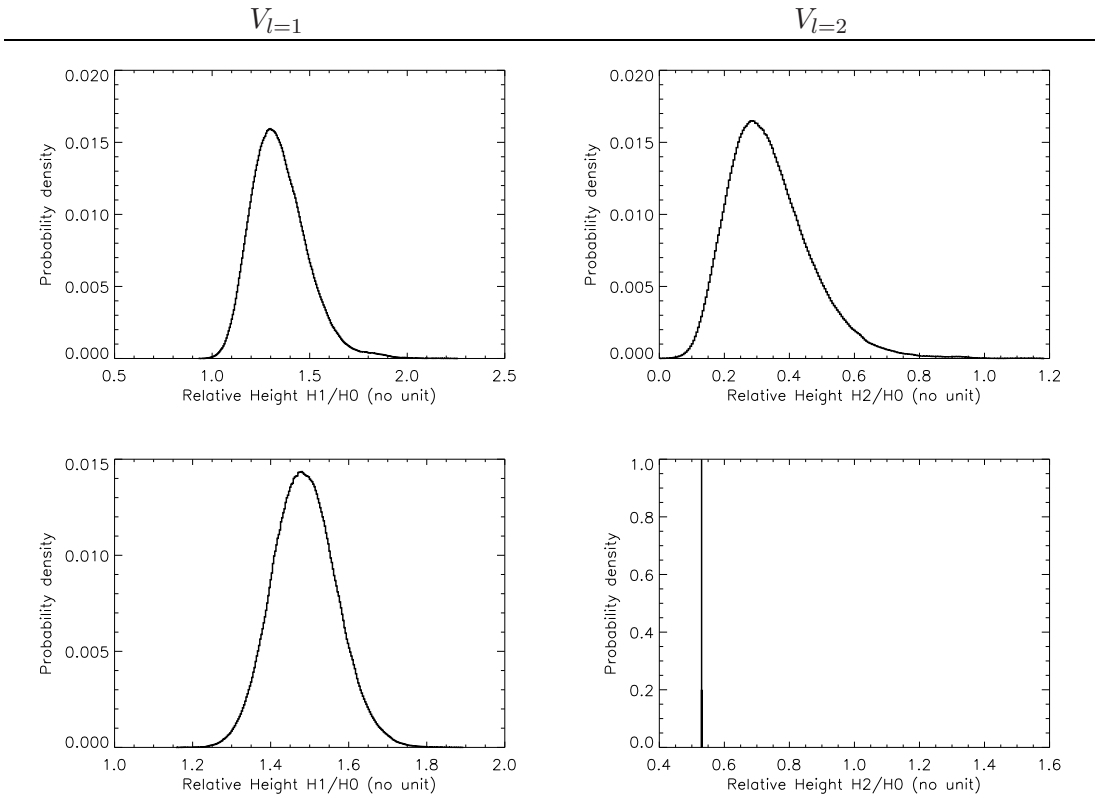


TABLE 4.12 – Hauteur relative du $l = 1$ et du $l = 2$ lorsque celles ci sont toutes deux libres ou lorsque celle du $l = 2$ est fixé à la valeur solaire. Dans ce dernier cas, la distribution est un dirac à la position $V_{l=2} = 0.53$. Fixer la hauteur relative du $l = 2$ fait remonter la valeur de la hauteur relative du $l = 1$ à une valeur très proche de la valeur solaire (médian pour $V_{l=1} = 1.49$). Ceci est une indication de plus en faveur de l'identification M_B^2.

solaire). Le collapsograme est généré en sommant le spectre après sa découpe en morceaux de longueurs égale à $\Delta\nu$. En somme, c'est une version collapsée du diagramme echelle. Seul les 12 ordres radiaux les plus intenses ont été collapsés. Au final, il est probable à 72% (±2%) que le $l = 3$ ne soit pas nécessaire pour décrire correctement le profil observé. Un examen visuel abonde d'ailleurs dans ce sens. Nous concluons donc qu'il n'est pas possible de distinguer le $l = 3$, malgré la remarquable qualité des données déjà atteinte.

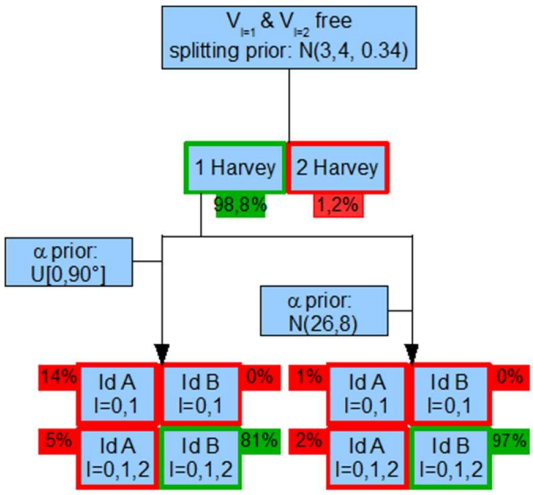


FIGURE 4.8 – Diagramme de décision obtenu lorsque l'on suppose *a priori* les hauteurs libres. Les modèles rejetés lors du processus décisionnel sont en rouge. Ceux acceptés en vert. Les valeurs des *a priori* fournis ici prévalent évidement face à celle de la Table 4.2. Initialement, on a cherché la possibilité de la nécessité d'un second profil de type Harvey pour décrire le bruit lors de l'ajustement des modes individuels. Cette hypothèse n'est jamais soutenue. Quelque soit l'*a priori* appliqué, la probabilité est toujours en faveur du modèle M_B^2. L'erreur sur chacune des probabilités est ici inférieure à 4%.

La grande majorité des valeurs de fréquences obtenues par les différents ajusteurs des données s'accorde à 1σ. Toutes s'accordent dans un intervalle à 1.5σ. Les barres d'erreurs sont extrêmement faibles dans l'intervalle $\nu \in [1300, 2000]$, particulièrement pour le $l = 1$ (Fig. 4.3.1) puisqu'il n'est peu ou pas affecté par la présence de l'indiscernable $l = 3$. Le $l = 2$ est en revanche très mal contraint. Les diagrammes échelles des données et de l'ajustement sont présentés Fig.4.3.1. Encore une fois, il n'est pas possible de visuellement établir l'identification. On pourra cependant faire remarquer que c'est la légère asymétrie provoquée sur l'excès d'énergie associé au couple $l = 0-2$ qui conforte les tests statistiques à choisir le modèle B. Pour s'en convaincre, on pourra regarder la Fig.4.3.1, représentant la tranche spectrale (sur une grande séparation) dans laquelle le rapport hauteur sur bruit est le plus élevé. Une vue d'ensemble du spectre est aussi fournie 4.3.1.

On trouvera les profils des hauteurs, largeurs et amplitudes en Fig.4.3.1. On observe globalement les mêmes tendances que pour le Soleil. Par exemple, les largeurs croissent avec l'ordre n. Il semble aussi y avoir un plateau autour du maximum d'amplitude. L'amplitude est remarquablement bien déterminé avec IRa-01+LRa-01, ce qui est bien moins le cas lors de l'analyse du IRa-01. C'est moins le cas pour les hauteurs et surtout pour les largeurs.

La Fig.4.3.1 montre la dépendance entre l'angle et le splitting. On remarque que tout deux sont corrélés. C'est un fait déja signalé dans Ballot et al. (2008) et nous aurons le loisir de l'illustrer

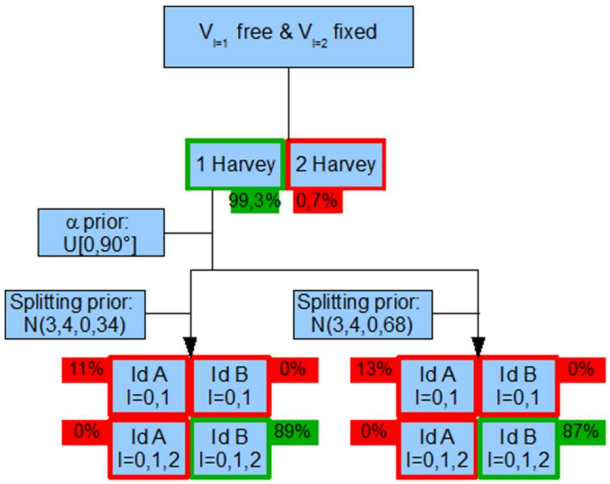


FIGURE 4.9 – Même chose que la Fig.4.8 lorsque l'on présuppose la hauteur relative du $l = 2$ fixée à la valeur solaire et celle du $l = 1$ libre. On notera que doubler la "tolérance" sur l'*a priori* du splitting ne change aucunement la probabilité. L'erreur sur chacune des probabilités des modèles est ici inférieure à 4%.

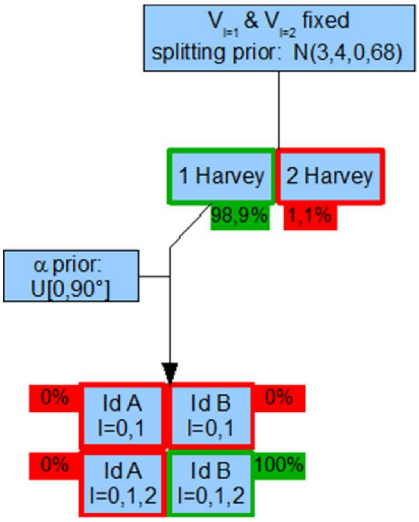


FIGURE 4.10 – Même chose que la Fig.4.8 lorsque l'on présuppose la hauteur relative du $l = 1$ et $l = 2$ fixées aux valeurs solaires. On fera remarquer que toutes les autres équipes ayant analysé les données (les ajusteurs) se sont placées dans ce cas. Nous n'avons pas jugé nécessaire de comparer différents scénarii sur les *a priori* ici car l'identification fournie est sans ambigüité possible ici. L'erreur sur chacune des probabilités est inférieure à 0.01%.

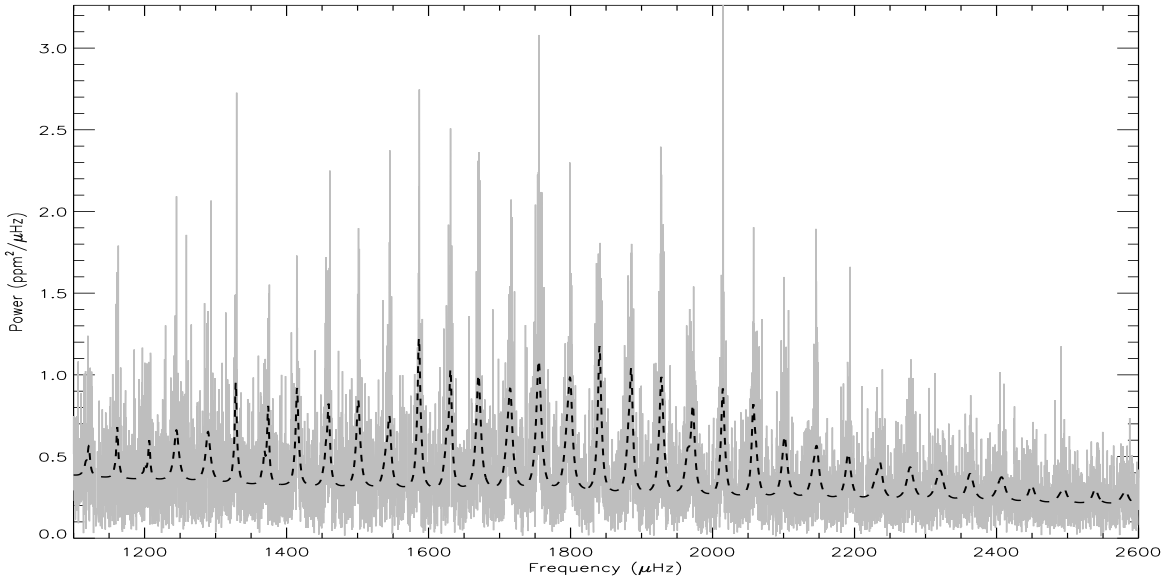


FIGURE 4.11 – Le spectre moyenné et l'ajustement associé.

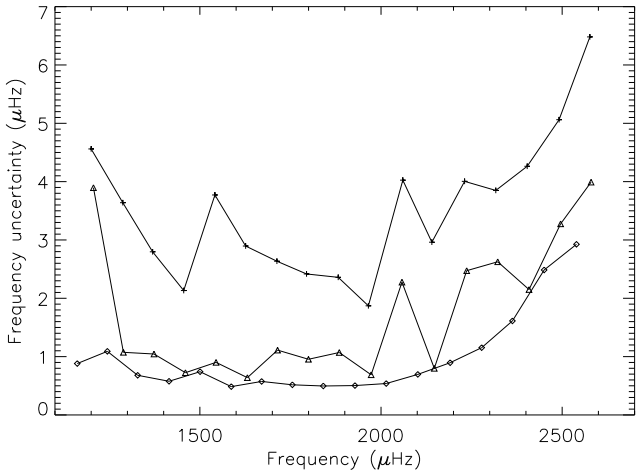


FIGURE 4.12 – Erreur à 1σ moyenne (borne supérieure et inférieure) pour les fréquences du $l = 0$ (triangles), $l = 1$ (losanges) et du $l = 2$ (croix) de HD49933 (IRa-01 + LRa-01) en fonction de la fréquence.

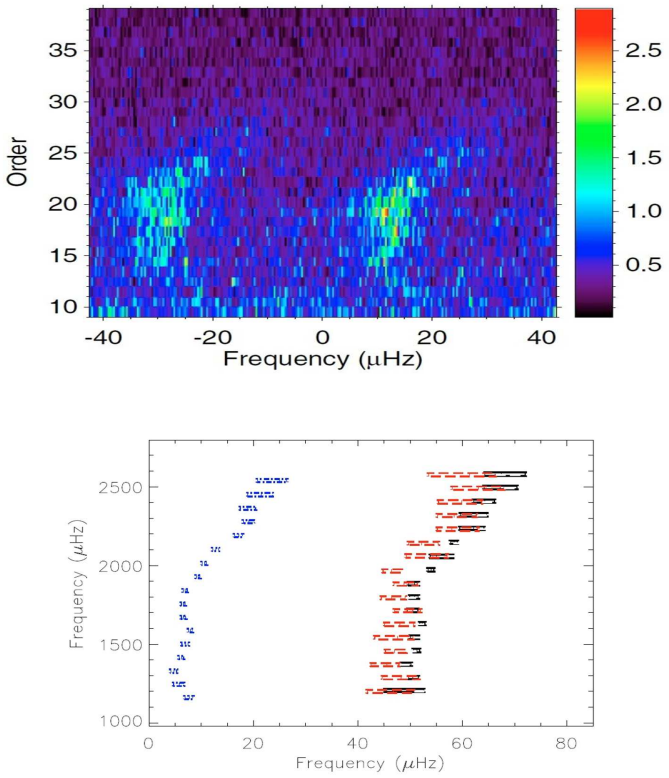


FIGURE 4.13 – Diagramme échelle du spectre de puissance moyen de HD49933 (IRa-01 + LRa-01) et des fréquences *a posteriori*.

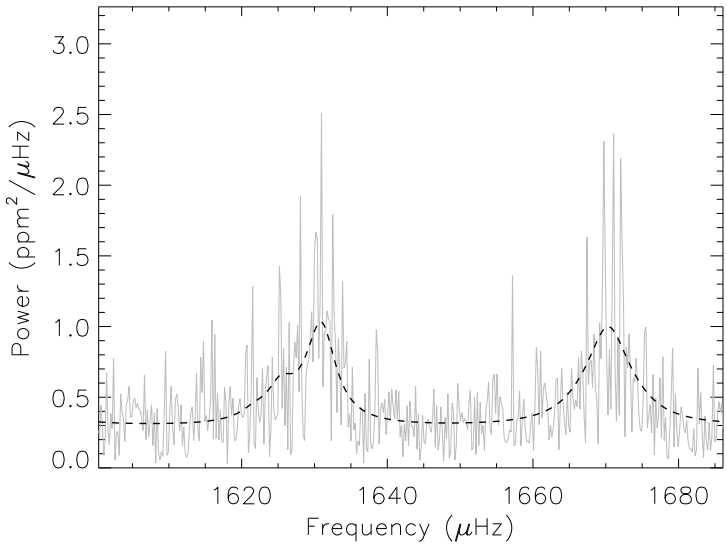


FIGURE 4.14 – Section du spectre de HD49933 pour laquelle le rapport hauteur sur bruit des modes est le plus élevé. On distingue assez clairement l'asymétrie sur l'excès d'énergie du coté du couple $l = 0 - 2$. Elle est moins évidente à discerner ailleurs dans le spectre.

un peu plus loin. Le maximum de densité de probabilité est obtenue autour de $\{22°, 3.9\,\mu\text{Hz}\}$. Sur la figure, la zone rouge et jaune représente assez bien l'erreur à 1σ. Avec cette identification ces valeurs sont cohérentes avec celles déduites du $v sin\, i$, du rayon et de la période rotationnel.

Alors que les analyses par MLE privilégiaient la mauvaise identification avec le IRa-01 (avec une probabilité de plus de 99.99%(!)), les analyses MCMC aboutissaient à une conclusion bien plus nuancée. L'utilisation d'une série temporelle plus longue et donc statistiquement plus fiable a permit de réconcilier ces deux approches. Les méthodes d'analyses classiques montrent leurs limites lorsque la fonction vraissemblance est fortement bruitée (*i.e.* constellé de maxima locaux). En conclusion, les méthodes bayésiennes sont lourdes mais bien plus efficaces que le MLE ou le MAP pour établir un jugement de la manière la plus juste possible.

Finalement, bien que HD49933 ait représenté un véritable défi à analyser, la richesse du spectre de cette étoile nous a permis d'extraire les propriétés individuelles de plus d'une cinquantaines de modes pulsations. L'observation par CoRoT de HD49933 a permis de mener un travail de pionnier en astérosismologie puisqu'aucun astre, à l'exception du soleil, n'avait été observé avec une telle précision. C'est sans conteste que nous pouvons affirmer que toute l'expérience acquise sur cette étoile a servi et servira l'analyse des autres étoiles de CoRoT et de Kepler.

4.3.2 Un autre cas difficile : HD181420

Tout comme HD49933, cette étoile est un cas assez difficile car même si nous disposons d'une série temporelle relativement longue (environ 150 jours), le rapport hauteur sur bruit est plus faible que pour HD49933.

HD181420 est une étoile en apparence assez comparable à HD49933, de type spectrale F2, de température de surface $T_{eff} = 6580 \pm 105K$ de masse $M/M_\odot = 1.311 \pm 0.063$ et de rayon $R/R_\odot = 1.595 \pm 0.032$. Sa luminosité est $log(L/L_\odot) = 0.63 \pm 0.03$ et de métallicité similaire au soleil : $[Fe/H] = 0.00 \pm 0.06$ selon Bruntt (2009) (l'indice $_\odot$ se réfère au Soleil). Une analyse a déjà été publiée dans Barban et al. (2009) dont voici les principales conclusions. HD181420 montre un clair excès de puissance imputé aux modes, autour de $1.5mHz$. Sa grande séparation moyenne est d'environ $75\mu Hz$. La rotation induit une modulation de la luminosité du fait du passage périodique de macro structure de surface. Ce pic se situe autour de $4.5\mu Hz$ et son étalement important (de 3 à 6 μHz) suggère une rotation différentielle prononcée. Ce pic fournit aussi une indication sur le splitting rotationnel des modes et sur l'angle d'inclinaison stellaire. En effet sur la base de considérations géométriques on a,

$$i = arcsin\left(\frac{v\, sin\, i}{2\pi R \nu_s}\right) \simeq 35°. \tag{4.10}$$

A l'issue de l'analyse menée par Barban et al. (2009), il n'a pas été possible de conclure quant à l'identification des modes. Un autre article de Mosser et al. (2009a) propose une méthode par autocorrelation pour identifier les modes. Il associe au pic à 1683 μHz un $l = 1$. C'est une méthode intéréssante car très rapide d'execution contrairement au calcul bayésien de probabilité. Toutefois nous devons rester prudent avec cette conclusion car la méthode est purement empirique. Une validation sur un ensemble important d'étoile doit être effectué pour identifier ses limites. Comme nous le verrons plus loin, nous trouvons la même identification des modes que Mosser & Appourchaux (2009). Notons aussi que la méthode proposée fournit aussi la bonne identification de HD49933, ce qui est encourageant.

Bien que de manière moins intensive que pour HD49933, une série de d'ajustements a été effectuée pour HD181420 en faisant varier quelques *a priori*. Nous allons tout d'abord présenter

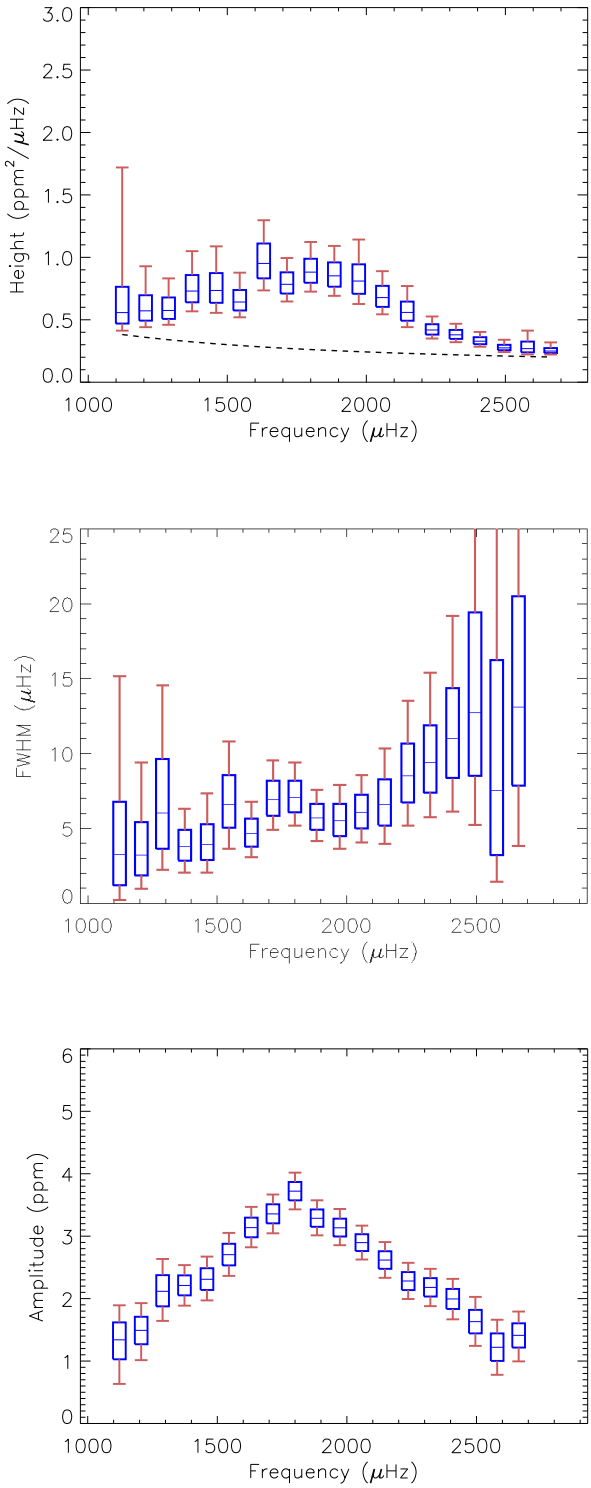


FIGURE 4.15 – Les profils de hauteur, largeur et d'amplitudes des modes pour HD49933 après analyse du spectre moyenné (IRa-01 + LRa-01).

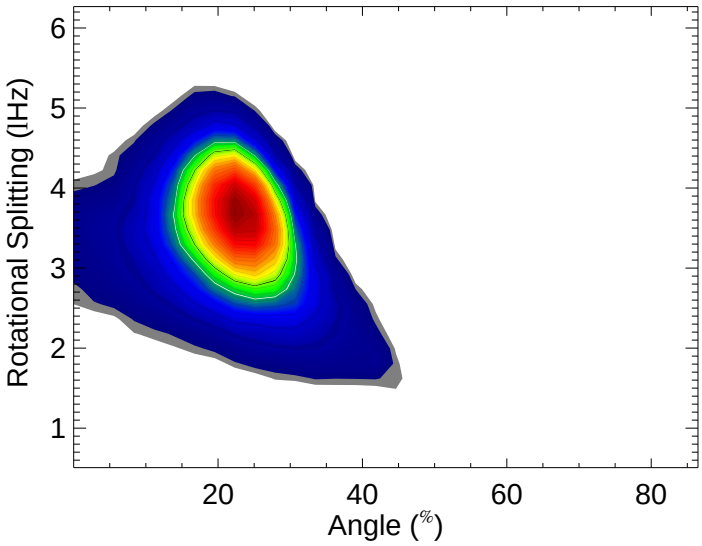


FIGURE 4.16 – Carte des dépendances entre angle d'inclinaison stellaire et splitting rotationnel pour HD49933 (IRa-01 + LRa-01).

les résultats importants sur les probabilités de modèles. Puis nous nous focaliserons sur les paramètres individuels et discuterons de leur dépendances respectives.

Identification des modes

Les *a priori* sur les hauteurs relatives étant les plus influents, nous avons testé 3 types de contrainte possibles, tout comme pour HD49933 (IRa-01 + LRa-01) :

- hauteur relative des modes libres (Scénario A).
- hauteur relative du l=2 fixée à la valeur solaire (Scénario B).
- hauteurs relative des modes fixée à la valeur solaire (Scénario C).

La rotation de surface semble assez mal contrainte, ce qui a justifié l'utilisation d'un *a priori* peu contraignant sur le splitting rotationnel (cf. Table 4.2). De la même manière que pour HD49933, nous avons constaté qu'un seul profil de Harvey suffisait statistiquement pour décrire la dépendance fréquentielle du bruit. Les Tables 4.13, 4.14, 4.15 résument tous les rapports concurrentiels et probabilités dans les différents scénarii envisagés. Le nombre d'échantillons acquis est suffisamment élevé (de 3 millions à 5.5 millions) pour garantir une erreur inférieure à 0.25 dex sur chaque log-probabilité. Remarquons que pour cette étoile le nombre d'échantillons nécessaire est très supérieur à celui nécessaire pour HD49933 pour atteindre la même précision (d'un facteur 2 environ). Sur la base de l'expérience d'ajustement MCMC que nous avons acquis, il y a deux circonstances qui peuvent expliquer cela.

Lorsque le modèle est inadéquat ou le niveau de bruit est élevé, les fdp (des fréquences notamment) s'éloignent de plus en plus de la gaussienne (multiples maxima locaux essentiellement). Dans ce cas, comme expliqué précédemment, l'efficacité d'échantillonnage décroit puisque celle-ci n'est optimale que si les fonctions à échantillonner et la loi de proposition sont de nature identiques [4]. Deuxièmement, le nombre de paramètres ajustés est légèrement supérieur au cas de

4. Se référer au Chap.3.2.2 pour plus de précision.

$M_k^l \backslash M_{k'}^{l'}$	M_A^1	M_B^1	M_A^2	M_B^2	$P_R(M_k^l \| y, I)$
M_A^1	1	16.44	73.70	5.47	79.6%
M_B^1	0.06	1	4.48	0.33	4.8%
M_A^2	0.01	0.22	1	0.07	1.1%
M_B^2	0.18	3.00	13.46	1	14.5%

TABLE 4.13 – Rapports concurrentiels et estimation des probabilités lors de l'analyse de HD181420 lorsque l'on suppose les hauteurs relatives $V_{l=1}$ et $V_{l=2}$ comme paramètres libres.

$M_k^l \backslash M_{k'}^{l'}$	M_A^1	M_B^1	M_A^2	M_B^2	$P_R(M_k^l \| y, I)$
M_A^1	1	22.20	67.43	2.52	68.7%
M_B^1	0.05	1	3.04	0.11	3.1%
M_A^2	0.01	0.33	1	0.04	1.0%
M_B^2	0.40	8.81	26.75	1	27.2%

TABLE 4.14 – Rapports concurrentiels et estimation des probabilités lors de l'analyse de HD181420 lorsque l'on suppose la hauteur relative $V_{l=2}$ comme paramètre fixé à la valeur solaire ($V_{l=1}$ étant libre).

$M_k^l \backslash M_{k'}^{l'}$	M_A^1	M_B^1	M_A^2	M_B^2	$P_R(M_k^l \| y, I)$
M_A^1	1	0.01	$1.60\,10^4$	0.00	0.1%
M_B^1	$1.8\,10^2$	1	$2.9\,10^6$	0.15	12.7%
M_A^2	0.00	0.00	1	0.00	0.0%
M_B^2	$1.2\,10^3$	6.90	$2.0\,10^7$	1	87.2%

TABLE 4.15 – Rapports concurrentiels et estimation des probabilités lors de l'analyse de HD181420 lorsque l'on suppose les hauteurs relatives $V_{l=1}$ et $V_{l=2}$ comme paramètres fixés à la valeur solaire.

HD49933 car nous avons un ordre radiale supplémentaire à ajuster.

Après observation de ces tables, nous pouvons constater qu'au premier regard il n'est pas possible d'établir avec certitude l'identification des modes car les résultats varient en fonction des *a priori* utilisés. Après une observation attentive, il est toutefois possible de faire les constatations suivantes :

1. Les scénarii A et B privilégient le modèle M_A^1 c'est à dire sans $l = 2$. Le rapport concurrentiel entre les deux modèles les plus probables (*i.e.,* M_A^1 *et* M_B^2) décroit toutefois de manière importante (la log-différence est divisée par deux) lorsque la hauteur relative du $l = 2$ est fixé à la valeur solaire. En marginalisant par le nombre de degrés, on a
   - Scénario A :

$$P_R(M_A|D,I) = P_R(M_A^1|D,I) + P_R(M_A^2|D,I) \simeq 80.7\% \quad (4.11)$$
$$P_R(M_B|D,I) = P_R(M_B^1|D,I) + P_R(M_B^2|D,I) \simeq 19.2\% \quad (4.12)$$

   - Scénario B :

$$P_R(M_A|D,I) = P_R(M_A^1|D,I) + P_R(M_A^2|D,I) \simeq 69.7\% \quad (4.13)$$
$$P_R(M_B|D,I) = P_R(M_B^1|D,I) + P_R(M_B^2|D,I) \simeq 30.3\% \quad (4.14)$$

2. Le scénario C privilégie fortement l'identification B avec $l = 2$ (M_B^2). En marginalisant, on obtient ainsi l'identification B à plus de 99.9%.

3. Si on suppose comme très improbable l'absence du degrée $l = 2$ (ce qui revient à imposer aux *a priori* sur les modèles M_A^1 et M_B^1 les valeurs $P(M_A^1|I) = P(M_B^1|I) = 0$, cf. Eq.3.43) alors on obtient une probabilité qui est systématiquement et significativement en faveur de l'identification M_B^2 avec une probabilité supérieure lorsque l'on contraint les hauteurs relatives,

$$\text{Scenario A} : P_R(M_B|D,I) \simeq 93\% \tag{4.15}$$
$$\text{Scenario B} : P_R(M_B|D,I) \simeq 96.5\% \tag{4.16}$$
$$\text{Scenario C} : P_R(M_B|D,I) > 99.99\% \tag{4.17}$$

L'hypothèse suggérée au point 3 est tout à fait valide puisqu'on a affaire à des pulsateurs solaires. Si on suppose cette hypothèse vraie, alors les seules sources empêchant la détection formelle du $l = 2$ sont le niveau de bruit et le degré de superposition des modes de même parité. En effet, comme pour HD49933 et plus généralement pour les étoiles plus chaudes que le soleil, la largeur des modes est très importante [5]. La petite séparation attendue étant voisine du splitting rotationnel, nous nous retrouvons dans une configuration de paramètres stellaires défavorables (plus encore que pour HD49933) pour mettre en évidence le $l = 2$.

Comme nous l'avons déjà suggéré Chap.4.2.1, il est aussi possible d'identifier la présence du $l = 2$ indépendamment du calcul de probabilité et donc de vérifier la pertinence de l'hypothèse suggérée au point 3 : si il n'y a pas de $l = 2$, le profil de la fdp de la hauteur relative du $l = 2$ doit porter la signature de la statistique du bruit. La Fig.4.3.2 montre les deux fdp associées à ce paramètre. Pour l'identification M_A^2, nous retrouvons approximativement une loi exponentielle décroissante (ponctuée par des minima locaux) caractéristique du bruit, ce qui n'est pas le cas du modèle M_B^2.

Au regard de tous ces éléments, nous pouvons conclure que l'identification la plus probable est la B et incluant les modes $l = 2$. Cette identification est la même que celle fournie par Mosser & Appourchaux (2009).

Nous voyons ici toute les possibilités offertes par un échantillonnage. Les raisonnements qui peuvent être menés peuvent être beaucoup plus complexes qu'avec une plus simple approche.

Une autre conclusion s'impose : contrairement aux simulations, il n'est pas possible de discriminer correctement entre modèles de degrés de complexités différents (*i.e.* de nombre de paramètres différents) lorsque les *a priori* sur la hauteur relative n'est pas fortement contrainte. Au Chap.4.2.2 nous disions que les simulations testaient la pertinence d'une méthode en supposant le modèle *exact*. Nous pensons donc que la discrimination entre modèles de tailles différentes échoue dans le cas présent parce que le modèle ajusté n'est qu'approximativement vrai.

Paramètres individuelles et dépendances

Dans cette section nous présentons les résultats principaux portant sur les paramètres *a posteriori* obtenus pour HD181420. Nous comparerons notamment le comportement des modèles M_A^2 et M_B^2.

Dans la Table de Figure 4.16 on a représenté les fdp et les cartes de corrélation pour le splitting et l'angle d'inclinaison stellaire en fonction des différents *a priori* imposés sur la hauteur relative des modes de degrés $l = 1$ et $l = 2$. On remarquera premièrement que le splitting est mal contraint et fortement multimodal. Il est possible que ceci soit dû à une mauvaise paramétrisation. Les hypothèses les plus probables sont selon nous :

5. Grâce aux données Kepler, il est même possible d'établir une relation empirique entre température et largeur des modes. Cette dernière croit selon une loi de puissance en fonction de la température effective de surface de l'étoile.

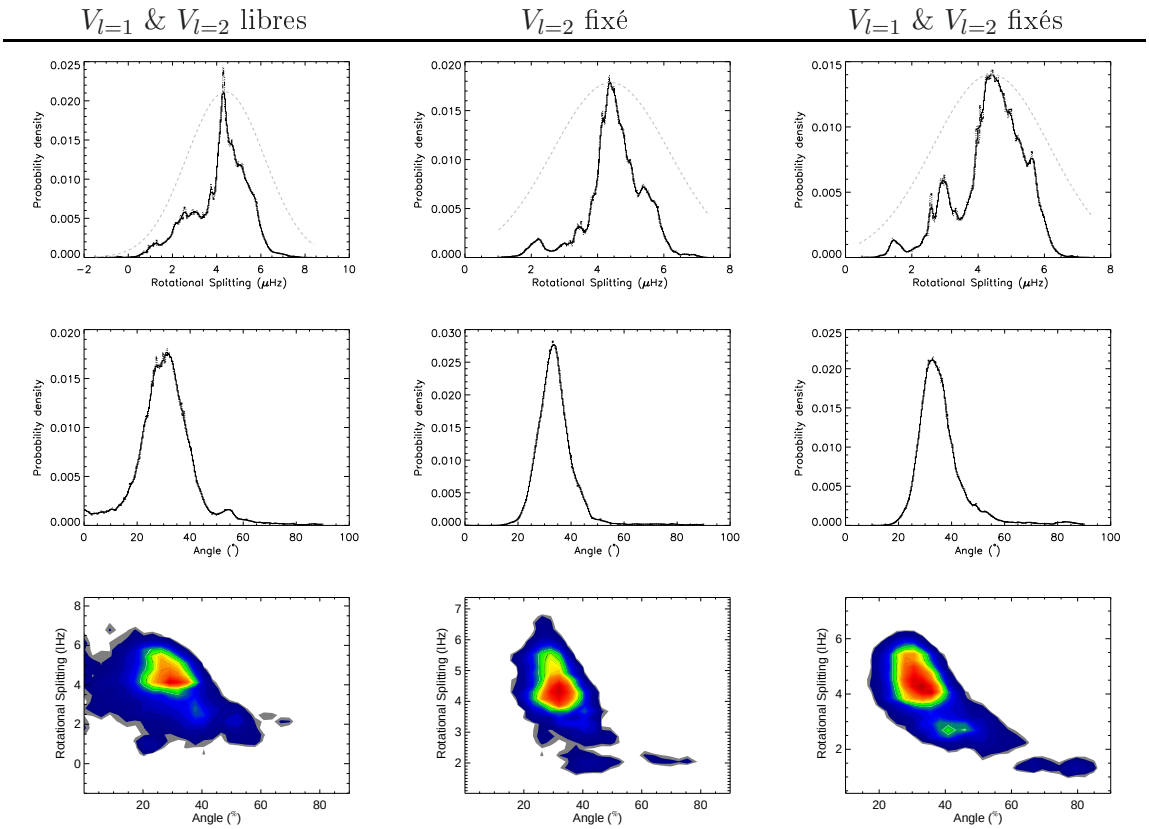


TABLE 4.16 – Fonction de densité de probabilité et carte de corrélation pour l'angle et le splitting. La distribution marginale obtenue pour le splitting est fortement chahutée quelque soit l'*a priori* utilisé. L'angle d'inclinaison est beaucoup plus stable. Ceci suggère que le splitting rotationnel est mal modélisé (symétrie des composantes rotationnelles et/ou variation de sa valeur en fonction de la fréquence). Il existe une forte corrélation entre angle et splitting.

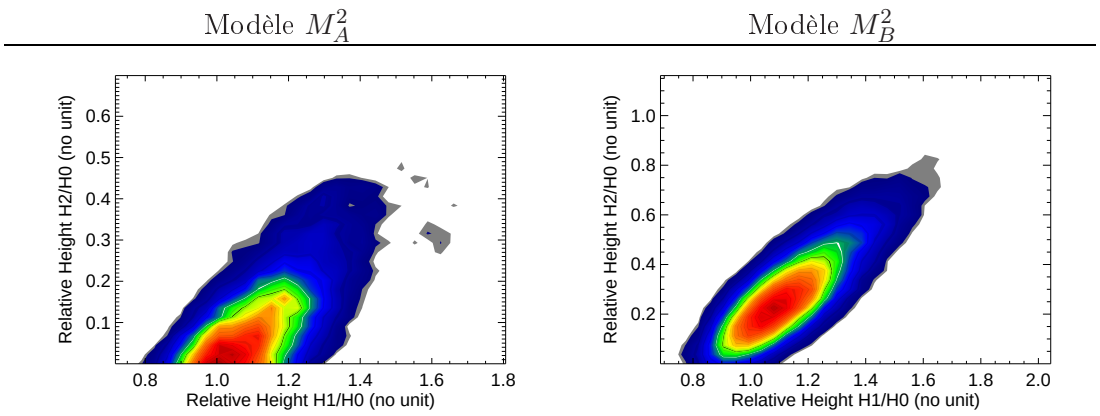


TABLE 4.17 – Carte de corrélation entre les hauteurs relatives des modes pour les scénarii M_A^2 et M_B^2. Lorsque la hauteur relative du $l = 1$ croît, la hauteur relative du $l = 2$ croit en proportion. Ce résultat est totalement indépendant de l'identification ajustée. Le même comportement peut être mis en évidence sur les autres étoiles que nous avons ajustées.

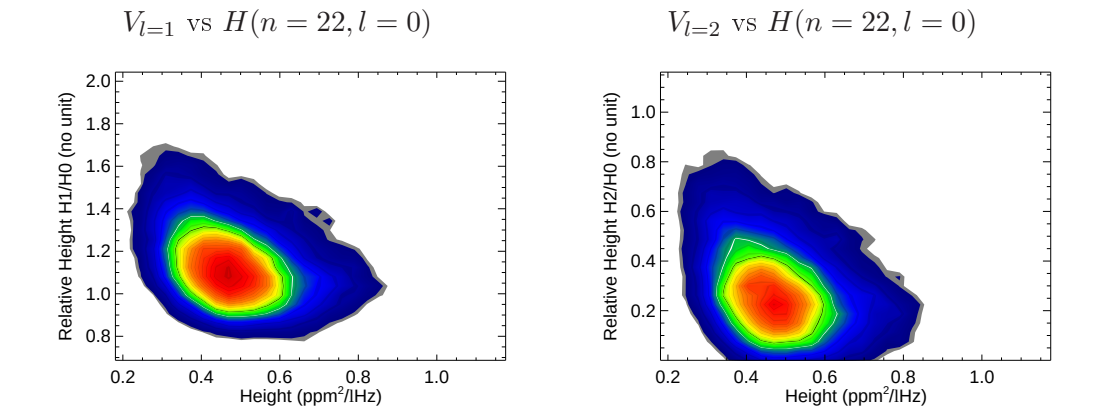


TABLE 4.18 – Carte de corrélation entre la hauteur relative des modes et l'une des hauteurs du mode $l = 0$ pour le modèle considéré comme le plus vraisemblable (M_B^2). Lorsque la hauteur du $l = 0$ augmente, la hauteur relative des modes $l = 1$ et $l = 2$ décroit en proportion. Ce comportement peut être mis en évidence sur tous les modes ajustés. Il est aussi commun à toutes les étoiles ajustées.

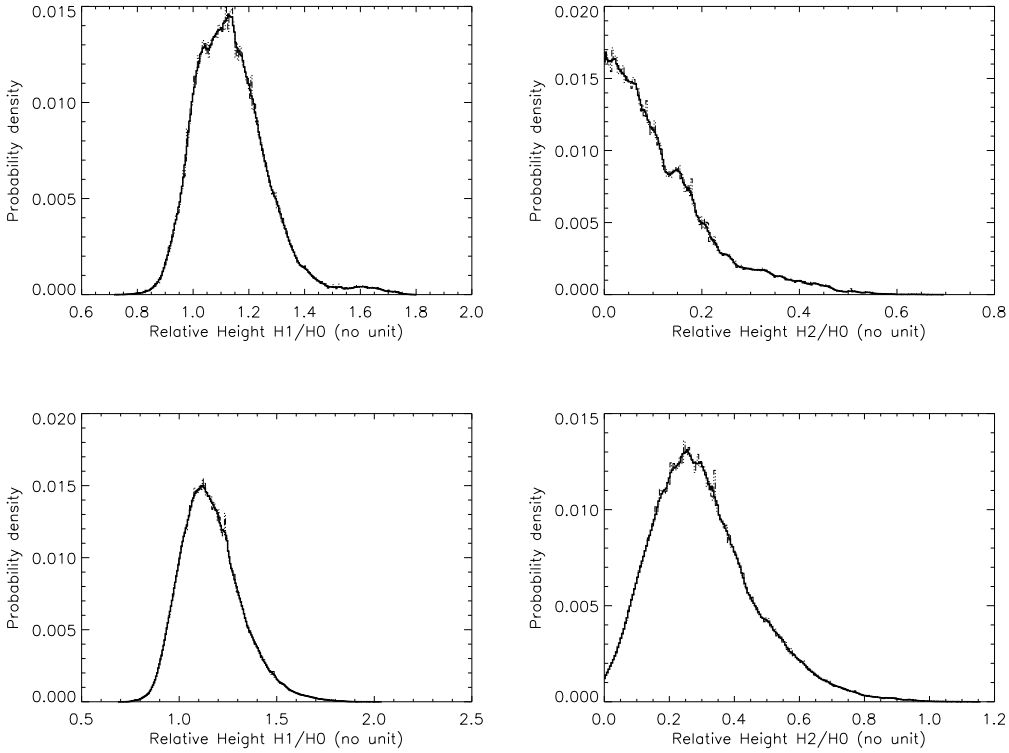


FIGURE 4.17 – Les hauteurs relatives des degrés $l = 1$ et $l = 2$ (scénario A) pour M_A^2 (haut) et M_B^2 (bas). La hauteur relative du $l = 2$ suit approximativement une loi décroissante.

1. il existe une asymétrie dans la position fréquentielle des composantes rotationnelles. le $m = 0$ n'est pas le barycentre des composantes $m = \pm j$.
2. l'hypothèse de rotation solide est incorrect. Le splitting varie fortement d'un mode à l'autre.
3. le niveau de bruit est trop élevé pour extraire une valeur fiable de ce paramètre.
4. le bruit ne suit pas exactement la statistique en χ^2 par endroit du spectre. En effet les pics orbitaux polluent significativement le spectre de cette étoile.

Les deux premières raisons possibles sont crédibles car nous savons que la rotation d'une étoile peut fortement influencer la position des modes m (le developement perturbatif d'ordre 1 est insufisant). Ceci est d'autant plus vraie que l'étoile tourne vite. Ballot (2010) a par exemple montré qu'avec une vitesse de rotation similaire à HD49933 (pour rappel, 3 à 4 jours de période), des effets d'asymétries (hypothèse 1) peuvent se faire sentir. On peut aussi imaginer une très forte rotation différentielle, similaire à celle constatée dans le Soleil, avec formation d'une tachoclyne (hypothèse 2).

Nous avons tenté d'évaluer l'éventualité 1 en supposant,

- une asymétrie des composantes rotationnelle constante (pour chaque l) et donc indépendante de la fréquence des modes ou,
- une asymétrie suivant une loi linéaire en fonction de la fréquence (pour chaque l).

Tous les modèles obtenus sont rejetés avec des rapports concurrentiels supérieures à la centaine (probabilité quasi-nulle). Précisons que le nombre de paramètre à ajouter varie de 3 à 6 en fonction des tests effectués, ce qui est peut-être trop important. Le spectre étant fortement bruité il est

fort possible que nous cherchions donc des effets trop fins à mettre en évidence en l'absence de connaissance *a priori* adéquate[6]. Le résultat obtenue ne signifie donc nullement qu'il n'y a pas d'asymétrie. C'est un problème classique en inférence : il est toujours plus facile de prouver qu'un phénomène physique est présent (en le mesurant) plutôt que l'inverse, l'absence de mesure concluante n'étant pas une preuve d'absence de phénomène. Selon nous, l'information contenue dans le spectre de puissance est tout simplement insuffisante pour établir avec certitude si il y a véritablement asymétrie des composantes rotationnelle. Autrement dit, nous ajoutons trop de paramètres et/ou ne limitons pas assez les solutions possibles par les *a priori*.

A l'avenir, il serait intéressant de rechercher une possible rotation différentielle en utilisant une description plus fine du splitting ; par exemple supposer une variation linéaire en fonction de la fréquence.

Quoiqu'il en soit, nous pensons que le niveau de bruit est trop élevé (hypothèse 3) et constitue très probablement la cause principale des difficultés auxquelles nous sommes confrontés pour comprendre le signal à notre disposition. Pour lever l'ambiguïté, ré-observer l'étoile serait donc la meilleure solution. On peut aussi envisager utiliser des *a priori* très contraignants pour rechercher les petites subtilités du signal. Il faut toutefois que ces connaissances *a priori* soient basées sur des considérations solides au risque d'aboutir à des conclusions erronées.

Une distribution moins étalée du splitting est obtenue à partir du moment où la hauteur relative du $l = 2$ est fixé. L'angle est bien contraint et relativement indépendant de la contrainte établie (l'erreur est toutefois plus importante si les hauteurs relatives ne sont pas fixées). La carte de corrélation montre sans surprise qu'il existe une forte corrélation entre angle d'inclinaison stellaire et splitting rotationnel. Ce résultat a déjà été souligné dans Barban et al. (2009) et Ballot et al. (2008). Ce qui est nouveau, c'est que le degré de corrélation semble dépendre de la connaissance *a priori* utilisé : la dispersion de la fonction de corrélation est moins importante lorsque les hauteurs sont libres. Une fois les hauteurs fixées, on retrouve le résultat classiquement admis : la ligne de crête de la fonction de corrélation entre splitting et angle d'inclinaison est une fonction décroissante de l'angle, de forme parabolique. Le même phénomène est visible pour HD49933 bien que de manière moins prononcée, probablement parce que les données sont de meilleure qualité.

La Table de Figure 4.17 montre la corrélation entre les hauteurs relatives pour les deux identifications possibles (modèles M_A^2 et M_B^2). On remarquera la dépendance très forte entre les hauteurs relatives des modes et ce, indépendamment de l'identification. Ce résultat est commun à toutes les étoiles ajustées jusque là (HD 49933, HD 181420, HD 52265[7] et 4 étoiles Kepler). Parallèlement, on observe une anti-corrélation entre hauteur du $l = 0$ et les hauteurs relatives du $l = 1$ et $l = 2$ (cf. Table de Figure 4.18).

Ces deux phénomènes s'expliquent assez facilement. De la théorie des oscillations stellaires, on peut montrer qu'il y a équi-partition de l'énergie entre les modes de même parité. Ainsi, on a en puissance,

$$\sum_{l\,impair} V_l = \sum_{l\,pair} V_l. \tag{4.18}$$

En supposant les hauteurs relatives pour $l > 2$ négligeables, on a donc $V_{l=1} \simeq V_{l=0} + V_{l=2}$. Augmenter la hauteur relative du $l = 2$ revient à diminuer la hauteur relative du $l = 0$. La dispersion que nous obtenons est probablement due à la propagation des incertitudes sur les hauteurs du $l = 0$ et aux approximations du modèle. Par exemple tous les modes non pris en compte (*i.e.* $l > 2$) participent *a fortiori* au bruit. Notons que cette relation n'a pas été utilisée

6. Nous avons utilisé des *a priori* de Jeffrey car nous ne disposons d'aucun renseignement précis sur l'importance de l'effet recherché.

7. Papier en cours de préparation par Jérome Ballot.

comme *a priori* mais elle pourrait éventuellement mieux contraindre l'ajustement et permettre une meilleure discrimination entre modèles.

Dans ce qui suit nous ne présenterons que les résultats obtenus lorsque la hauteur relative du $l = 2$ est fixée. En effet, il n'y a pas de différence majeures dans les paramètres individuels des modes entre ce scénario (précédemment nommé B) et le scénario dans lequel toutes les hauteurs relatives sont fixées (précédemment noté C). Ceci du fait de barres d'erreurs relativement importantes. Par contre, comme nous l'avons déjà expliqué, la superposition des modes de même parité peut modifier leur estimation et il est préférable de "décorreler" les degrés $l = 0$ et $l = 2$ mitoyens en utilisant au mieux l'information que nous jugeons pertinente, à savoir que la hauteur relative du $l = 2$ est très proche de la valeur solaire. C'est pourquoi nous ne présentons pas non plus le scénario A plus en détail.

On a représenté le spectre de puissance superposé au meilleur ajustement du modèle M_B^2 scénario B, Fig.4.18 et 4.19. On remarquera que malgré une richesse exceptionnelle du spectre en terme de nombre de modes (près de 60 identifiés), il est bien difficile de les extraire de manière précise : l'étalement des modes est très important du fait de l'influence conjointe d'une rotation de l'ordre de quelques jours (*i.e.* splitting élevé) et de temps de vie caractéristique des modes très courts (*i.e.* largeur des modes importantes).

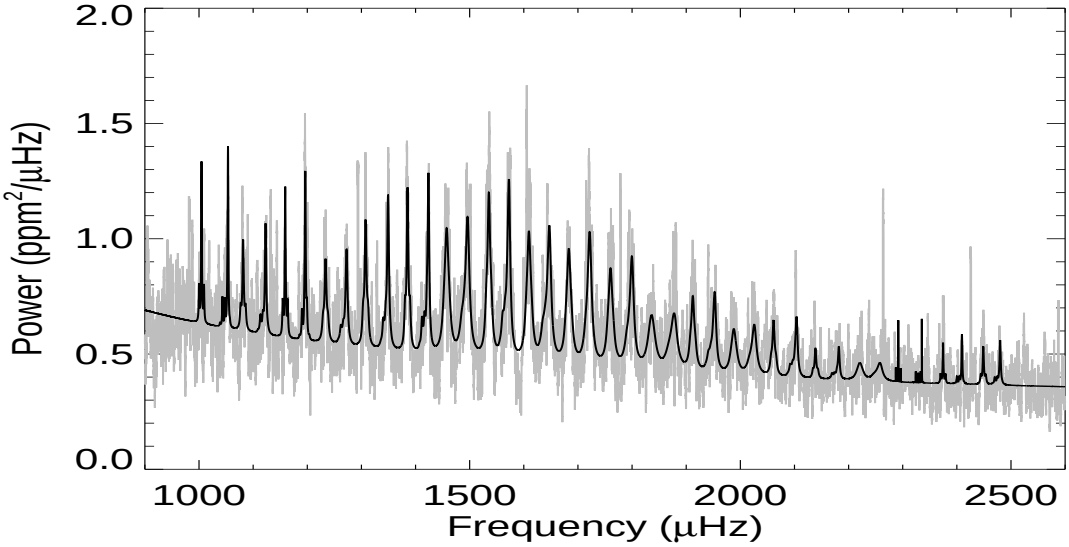


FIGURE 4.18 – Le spectre de puissance de HD181420 superposé au meilleur ajustement (modèle M_B^2). Le spectre est extrêmement riche avec plus 20 ordres radiaux.

La Fig.4.20 regroupe les profils en fonction de la fréquence des principaux paramètres des modes (hauteur, largeur, amplitude). On pourra encore une fois faire le parallèle avec les résultats obtenues pour HD49933. Bien que le profil des hauteurs soit relativement similaire à celui de HD49933 (profil grossièrement gaussien), des différences notables sont visibles pour les largeurs et pour les amplitudes : les largeurs des modes augmentent jusqu'à environ 1850 μHz pour diminuer ensuite. L'observation du profil d'amplitude révèle qu'il y a une discontinuité dans le profil d'amplitude autour de 1450 μHz (sixième ordre radial ajusté). Cette discontinuité semble coïncider avec une rupture de régularité de l'oscillation de l'ordre radial $l = 1$ (visible dans le diagramme échelle Fig.4.21). Il est donc possible que ce mode soit mixte et en ce sens, cela fournit

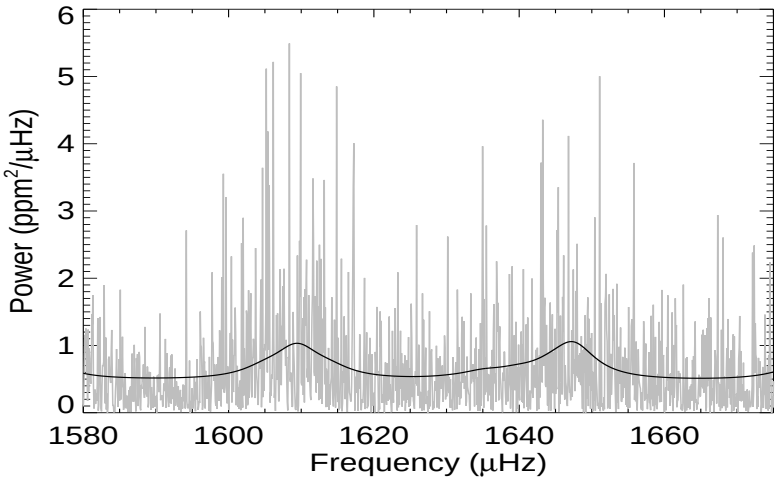


FIGURE 4.19 – Le spectre de puissance de HD181420 superposé au meilleur ajustement (modèle M_B^2) dans la région du maximum de hauteur des modes. Le signal est extrêmement bruité. A cause d'une valeur d'un splitting et d'une largeur des modes importante, l'énergie des modes est très étalée dans le spectre. L'extraction et l'identification des modes n'en est que rendue plus difficile.

une contrainte forte sur le stade évolutif de HD181420. Nous discuterons de cet éventualité dans le Chap.6.3 consacré au modèles de structure interne. Il est aussi possible que ce soit un effet des pics orbitaux. En effet, à cause de la largeur très importante des modes dans cette région, les $l = 1$ (situés autour de 1307 μHz, 1460 μHz et 1610 μHz) sont pollués par les pics orbitaux. Nous ne pensons pas cette hypothèse valide car les modes sont extrêmement large par rapport à la largeur typique des pics orbitaux (cf. Fig.4.22). De plus, nous ne tenons pas compte des points supposés affectés. Enfin on peut supposer que ce sont les hypothèses du modèle ajusté (splitting constant, etc...) qui sont en cause. Mais dans ce cas, les barres d'erreurs tiennent compte implicitement de cela. Ajoutons aussi que le rapport hauteur sur bruit est de l'ordre de 1. L'excitation stochastique est significative. La variance du bruit [8] est de l'ordre de 0.5 ppm^2/μHz et rares sont les points du spectre dépassant un seuil à 3σ. C'est pourquoi on rappelle qu'il est important de se fier à l'intervalle de confiance et *pas seulement* à la valeur médiane.

8. Pour une statistique du χ_2^2, la variance est égale à la moyenne.

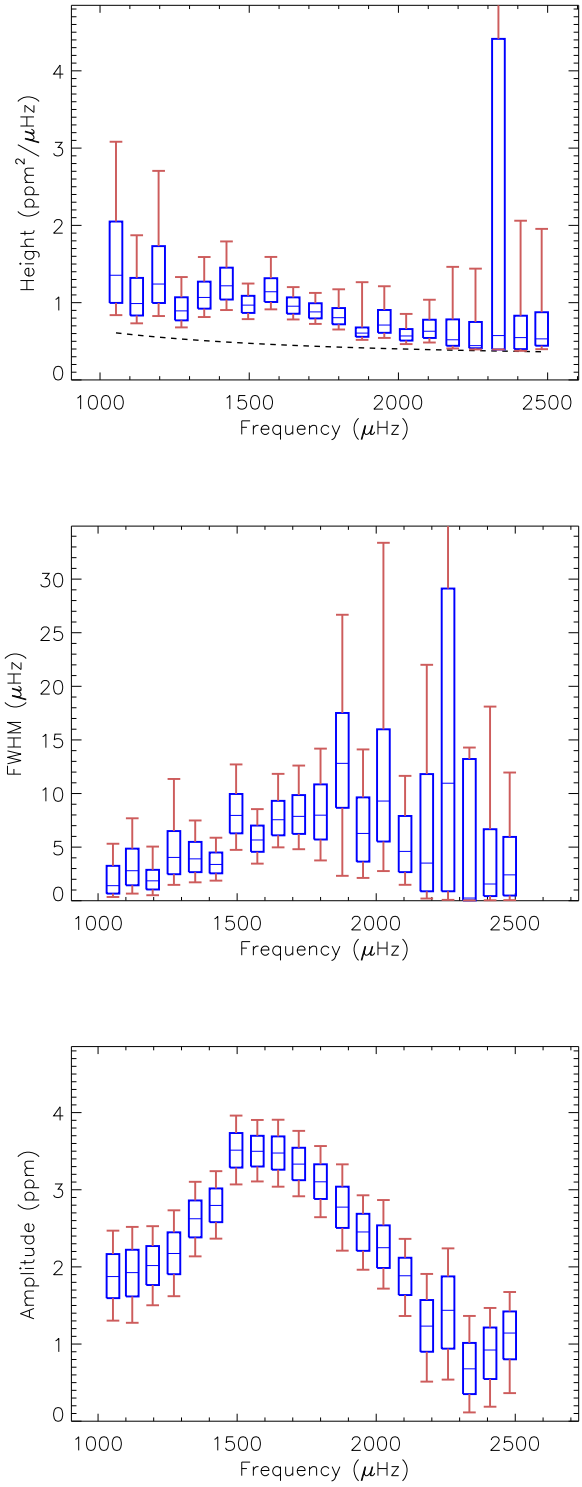


FIGURE 4.20 – Les profils de hauteurs, largeurs et d'amplitudes des modes pour HD181420. Les profils sont radicalement différents de ceux de HD49933 pour les largeurs et amplitudes. Il y a notamment une discontinuité dans le profil d'amplitude entre le 6ème et le 7ème ordre radial.

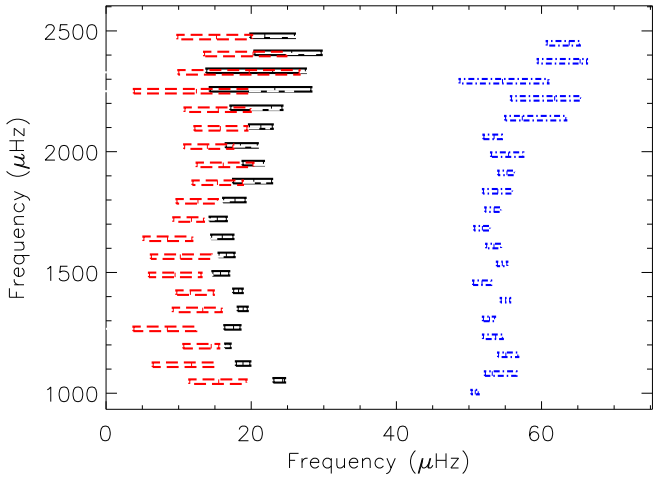


FIGURE 4.21 – Diagramme échelle de HD181420. Les erreurs sont particulièrement importantes. On remarquera que la position du 6ème $l = 1$ ($\nu = 1385.18$ μHz) du diagramme ne suit pas l'oscillation quasi-périodique visible jusqu'à 2000 μHz.

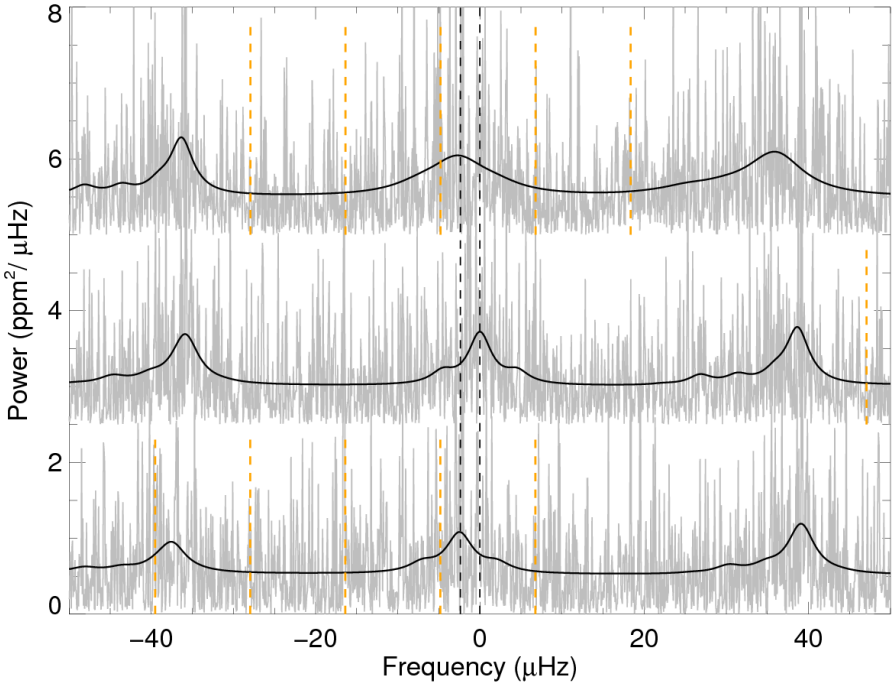


FIGURE 4.22 – superposition de 3 segments du spectre de puissance après re-centrage sur la position du mode $l = 1$ à 1385.18 μHz. De bas en haut, nous avons donc le $l = 1$ à 1307.69, 1385.18 et à 1457.46 μHz, dont les lignes verticales noires signalent les positions médianes. Celui à 1385.18 ne suit pas l'oscillation visible sur le diagramme échelle. Les lignes verticales en orange signalent la position attendue des pics orbitaux. La perturbation orbitale est très localisée. En générale elle n'affecte pas plus de 1 ou 2 points du spectre. Faisons remarquer que les points supposés affectés n'ont pas été pris en compte lors de l'ajustement (3 points par pics sont ignorés).

Chapitre 5

Extraction automatique et systématique de paramètres stellaires

Sommaire

5.1 Problématique 113
5.2 Extraction d'information générale sur les modes 114
5.2.1 Méthodologie 114
5.2.2 Les connaissances *a priori* utilisées 115
5.3 Extraction des modes 115
5.3.1 Comment automatiser l'ajustement des modes ? 116
5.3.2 Contraintes de continuité des fréquences 117
5.3.3 Contrainte de continuité des hauteurs et largeurs 120
5.3.4 Le bruit lors de l'ajustement des modes 123
5.3.5 Remarque sur l'angle d'inclinaison et le splitting rotationnel 123
5.3.6 Résumé de la méthode 124
5.4 Test sur les données solaires : mise en évidence du cycle solaire . . 125
5.5 Performance de l'algorithme sur un échantillon d'étoiles simulées . 130

5.1 Problématique

Nous avons vu que lorsque le rapport hauteur sur bruit était faible, nous pouvions être dans l'incapacité d'établir avec discernement quelle était l'identification des dits modes. De plus, ceci peut être accentué par des caractéristiques intrinsèques de l'étoile observée, fortement défavorable à une analyse correcte de certains paramètres stellaires (superposition des modes ou angle d'inclinaison quasi-nul par exemple). Dans ces cas, une analyse intensive mettant en oeuvre des méthodes lourdes d'échantillonnage à base de MCMC ont montrées leur efficacité. Dans un contexte où seuls quelques objets sont observés, tel que c'est le cas avec la mission CoRoT, ce type d'outils peut fournir des résultats dans des temps raisonnables. Cependant dans le cadre de la mission Kepler, une analyse par échantillonnage MCMC n'est pas envisageable pour l'instant. En effet comment analyser plusieurs milliers d'objets efficacement et rapidement lorsque le temps d'analyse pour un seul de ces objets nécessiterait plusieurs semaines de calcul sur une machine actuelle ?

Pour répondre à cette problématique, il est donc nécessaire d'élaborer des outils d'analyses statistique relativement efficaces et rapides. Une grande précision des résultats n'est alors pas le

seul critère à prendre en compte. Bien entendu, le gain en rapidité se fera au détriment de la qualité des paramètres estimés. C'est pourquoi il ne faut pas excessivement réduire la complexité des modèles mis en oeuvre (par exemple, supposer la hauteur des modes indépendante de la fréquence est une hypothèse trop forte).

Ce chapitre a pour objet d'essayer de répondre à cette problématique en présentant une approche globale d'analyse des modes stellaires pour les étoiles de type solaires dans leur séquence principale. Ce travail a vu le jour suite à mon séjour en Angleterre, à l'Université de Birmingham, sous la direction du Dr William Chaplin. Dans ce qui suit, nous exposerons les choix effectués concernant les moyens d'analyse, les hypothèses de travail et les performances de la méthode.

Tout d'abord, exposons le problème. Un réseau au sol ou un satellite observe un champ précis du ciel grâce à un photomètre. Dans ce champ, la plupart des étoiles ont déjà été cataloguées lors de campagnes antérieures et nous disposons la plupart du temps d'informations photométriques, astrométriques et spectroscopiques basiques. Les estimations couramment à notre disposition et sur lesquelles nous porterons notre intérêt sont :

- la température effective de surface de l'étoile : T_{eff}
- la luminosité de l'étoile : L.
- le logarithme de gravité en surface : $log(g)$. Gravité qui est proportionnelle au rapport M/R^2 où la M est la masse et R le rayon de l'étoile : $g = \frac{GM}{R^2}$ (G étant la constante gravitationnelle).

Les mesures effectuées par notre photomètre nous permettent de construire un spectre de puissance représentatif des pulsations de chacune des étoiles du champ.

Quelle est alors la manière la plus efficace d'extraire les paramètres individuels des modes stellaires automatiquement et sur la base de ces seules mesures ?

5.2 Extraction d'information générale sur les modes

5.2.1 Méthodologie

Comme le montrent l'Eq.1.52 et l'Eq.1.53, à partir du $log(g)$ et de T_{eff} il est possible d'estimer l'amplitude maximale des modes et leur position moyenne dans le spectre. Par la suite, en se basant sur l'Eq.1.54, la grande séparation peut aussi être estimée.

De plus, Chap. 4.3.1, nous avons montré qu'en première approximation, il était possible de modéliser l'amplitude ou la hauteur des modes de pulsation stellaire par une fonction gaussienne décrivant l'enveloppe des dit modes. Le bruit stellaire peut être quant à lui modélisé par un ensemble de fonction de type Harvey, pour lesquelles nous avons justifié l'usage au Chap. 1.2.1.

En pratique, les lois d'échelles définies dans le Chapitre introductif (Chap.1.2.8) sont très approximatives : elles ne décrivent correctement que le comportement moyen d'une large population d'étoile. Chaque étoile peut, de par ses particularités, être très éloignée de la tendance moyenne donnée par ces lois d'échelles. Qui plus est, nous ne disposons *a priori* d'aucune information sur le jeu de paramètres associés au bruit. L'information *a priori* à ce stade est donc très parcellaire. Dans ces conditions, le moyen le plus efficace d'établir un ajustement du spectre est d'utiliser un algorithme d'échantillonnage, dans le cas présent à base de MCMC.

Néanmoins, nous voulons limiter le temps de calcul. Pour allier à la fois qualité et rapidité nous avons choisi d'effectuer une première recherche en aveugle (échantillonnage MCMC) du meilleur jeu de paramètre pour le bruit et les modes décrits de manière globale (ν_{env}, H_{env}, σ_{env}). Mais cette recherche rapide ne sert qu'à approcher la solution optimale. En effet, un

algorithme d'échantillonnage MCMC est très efficace pour s'approcher de la solution la plus probable et ce, indépendamment des paramètres initiaux.

Une fois la solution approchée, typiquement au bout de 15 000 itérations [1], une approche MAP prend le relais pour affiner le résultat et calculer les barres d'erreurs associées grâce à une estimation de la matrice hessienne.

Le temps total de convergence (MCMC + MAP) vers la solution est relativement court (de l'ordre de 3 min par étoile sur un coeur de processeur Intel i7 cadencé à 2.67Ghz) car le nombre de paramètres est relativement faible (inférieur à la dizaine).

Un second problème consiste à établir si l'excès d'énergie attribué aux modes est suffisamment significatif pour permettre une analyse plus approfondie des modes (l'extraction des paramètres individuels). Pour cela, on compare deux modèles, l'un n'ayant que les composantes de Harvey et du bruit blanc (modèle sans mode, à 7 paramètres, Eq.5.1a), l'autre ayant à la fois ces composantes et une gaussienne (modèle avec mode, à 10 paramètres, Eq.5.1b). Cette comparaison se fait en calculant le rapport concurrentiel entre les deux modèles testés pour chaque étoile, tel qu'expliqué au Chap.2.2.3 (Eq.2.19). On a,

$$M_1(\nu) = \sum_{i=1}^{n=2} \frac{A_i}{1 + (B_i\nu)^{p_i}} + C. \tag{5.1a}$$

$$M_2(\nu) = M_1(\nu) + H_{env} e^{-0.5(\nu-\nu_{env})^2/\sigma_{env}^2}. \tag{5.1b}$$

Le seuil de réjection entre signal utile et bruit est fixé à 50%. Autrement dit,

- si $O_{M_2,M_1} > 1$ (*i.e.* $P(M_2|y, I) > 50\%$) alors l'étoile est retenue pour une analyse plus fine,
- sinon, alors l'étoile est rejetée : aucune analyse supplémentaire n'est effectuée.

Pour assurer la stabilité des résultats sur l'ensemble relativement hétérogène d'étoiles observées, nous avons préféré limiter le nombre de profils de type Harvey à 2. Nous négligeons le profil dominant à très basse fréquence caractérisant les phénomènes très lents en surface de l'étoile, tel que les taches stellaires. Ces phénomènes se manifestent à très basse fréquence. On restreint donc l'ajustement à la plage $[80 - F_N]$ μHz (F_N étant la fréquence de Nyquist).

5.2.2 Les connaissances *a priori* utilisées

Les *a priori* utilisés ici sont plutôt subjectifs car avec quelques dizaines de milliers d'échantillons seulement, l'algorithme MCMC ne fournit qu'un jeu de paramètres proche du maximum, sans estimer correctement les erreurs. Obtenir une estimation correcte de la moyenne nécessite moins d'échantillons que le calcul de la variance et des moments supérieurs. Le tableau 5.1 synthétise les informations *a priori* utilisées. Pour résumer, à ce stade nous faisons confiance à la valeur fournie par l'algorithme MCMC jusqu'à un certain degré, défini de manière empirique et résumé dans la table précitée. Par exemple, on estime que la hauteur de l'enveloppe ne peut excéder de 50% la valeur moyenne trouvée à partir de l'échantillonage.

5.3 Extraction des modes

A ce stade, pour les étoiles ayant un excès de puissance imputé aux modes suffisamment significatif (comparaison entre les modèles M_1 et M_2), on dispose d'une information globale sur les modes des étoiles. Il s'agit donc maintenant d'établir une procédure automatique qui :

1. On pourra faire le parallèle avec les simulations présentées Chap.4.2.1 dont le nombre de paramètres est similaire. Lors de ces simulations il a fallu une centaine de milliers d'échantillons pour reconstruire précisément les fonctions de densité de probabilité. Ici, ce n'est pas l'objectif recherché.

TABLE 5.1 – Synthèse des *a priori* utilisés lors de la phase de recherche de l'excès de puissance associé au modes de pulsation. Les indices 0 se rapportent aux résultats obtenus grâce à l'algorithme de recherche des paramètres initiaux (algorithme MCMC). y correspond au spectre de puissance. On retrouvera l'ensemble des figures et significations des variables en Annexe A.

paramètre	*a priori*	unité
H_{env}	$U(0,\ 1.5H_{env,0})$	$ppm^2/\mu Hz$
ν_{env}	$GUG(\nu_{env,0} - 2\sigma_{env,0},\ \nu_{env,0} + 2\sigma_{env,0},\ 0.5\sigma_{env,0},\ 0.5\sigma_{env,0})$	μHz
σ_{env}	$GUG(0.6\sigma_{env,0},\ 1.4\sigma_{env,0}, 0.1\sigma_{env,0},\ 0.1\sigma_{env,0})$	μHz
A_1	$U(0,\ max(y))$	$ppm^2/\mu Hz$
B_1	$U(10,\ 5000)$	$ksec$
p_1	$p_1 = 2$ (fixé)	/
A_2	$U(0,\ max(y)/4)$	$ppm^2/\mu Hz$
B_2	$U(0,\ 56)$	$ksec$
p_2	$U(0.5,\ 5)$	/

1. Détermine le nombre de modes à considérer lors de l'ajustement
2. Identifie individuellement les modes
3. Détermine les paramètres initiaux à utiliser
4. Effectue l'ajustement.

5.3.1 Comment automatiser l'ajustement des modes ?

Nous avons vu au Chap. 1.2.5 que dans le spectre de puissance, la répartition en fréquence des modes des étoiles de type solaire satisfait approximativement la loi asymptotique (Eq. 1.28). Rappelons la :

$$\nu(n,l) \approx (n + l/2 + \epsilon)\Delta\nu - l(l+1)D_0. \tag{5.2}$$

Les modes de même degré ont une distribution régulière, dont la distance moyenne n'est autre que la grande séparation $\Delta\nu$ et dont le point d'origine est caractérisé par D_0 et ϵ. Si $\Delta\nu$, ϵ et D_0 sont estimés, cette loi peut donc servir de point de départ pour la construction d'un modèle initial à ajuster.

Il existe diverses approches pour estimer $\Delta\nu$ automatiquement. On peut citer par exemple l'approche utilisée par Mosser & Appourchaux (2009), basée sur l'autocorrelation de la série temporelle ou encore l'approche présentée par Mathur et al. (2010) basée sur le spectre du spectre de puissance. Durant la thèse nous n'avons pas développé notre propre outil d'extraction automatique d'une estimation de la grande séparation car nous nous sommes reposés sur des résultats déjà existants lorsque cela a été nécessaire (*i.e.* pour l'analyse des étoiles Kepler). Par la suite, on nommera $\Delta\nu^*$ la valeur estimée *a priori* de la grande séparation.

D_0 est assez facile à estimer au premier abord car empiriquement, il a été montré qu'il valait une petite fraction de la grande séparation,

$$D_0 \approx \frac{2}{100}\Delta\nu. \tag{5.3}$$

Connaissant la grande séparation et D_0, il est alors possible d'estimer ϵ. Pour cela il nous faut tout de même une information supplémentaire : au moins une fréquence d'un mode $\nu(n,l)$

doit être connue. La stratégie adoptée consiste à "récupérer" le point de plus haut rapport signal à bruit dans la région où les modes sont observés. On appellera ν^* la fréquence associée à ce pic. Typiquement l'extraction se fait sur un intervalle de $3\sigma_{env}$ autour de ν_{env}.

Le pic extrait peut alors être attribué soit à un $l = 0$, soit à un $l = 1$ mais nous n'avons aucun moyen *a priori* de connaitre son degré l. Cependant, nous savons que pour intervertir le $l = 0$ et le $l = 1$, il suffit d'ajouter/retrancher approximativement $\frac{1}{2}$ à ϵ. En effet, grâce à l'Eq.5.2 et l'Eq.5.3, on peut écrire,

$$\epsilon^* \approx \frac{\nu^*}{\Delta\nu*} - n^* - \frac{l}{2} + 0.02l(l+1) \tag{5.4}$$

avec n^* la partie entière de la fraction $\nu^*/\Delta\nu^*$:

$$n^* = \left\lfloor \frac{\nu^*}{\Delta\nu^*} \right\rfloor . \tag{5.5}$$

Nous sommes encore une fois confrontés au problème d'identification des modes de degré pair et impair. Il nous faudra donc encore une fois tester les deux éventualités possibles pour les excès d'énergie avérées des modes. Cette comparaison se fera en calculant le rapport concurrentiel dans un contexte MAP. Nous aurons l'occasion d'y revenir.

Il est utile d'écrire la largeur σ_{env} de l'enveloppe des modes proportionnellement à la grande séparation. Dans ce cas, l'ordre radial minimum et maximum peut être estimé grâce à la relation,

$$n_{\pm} = \frac{\nu_{env} \pm \alpha\sigma_{env}}{\Delta\nu^*}. \tag{5.6}$$

α doit être ajusté pour tenir compte de suffisamment de modes, mais il est inutile qu'il soit trop grand car alors, les modes les plus éloignés de ν_{env} ont des hauteurs quasi nulles et sont totalement noyés dans le bruit. Utilisant une procédure empirique (essai/erreur), nous avons remarqué que $\alpha = 2.5$ est une valeur adéquate dans la plupart des cas étudiés durant la thèse.

Enfin, le nombre de modes étant entier, nous avons décidé de prendre les quantités arrondies par défaut $\lfloor n_- \rfloor$ et arrondies par excès $\lceil n_+ \rceil$. On peut alors définir un équivalent fréquence de ces grandeurs, $\check{\nu} = \lfloor n_- \rfloor \Delta\nu^*$ et $\hat{\nu} = \lceil n_+ \rceil \Delta\nu^*$.

5.3.2 Contraintes de continuité des fréquences

Modélisation des fréquences

La loi asymptotique peut être vue comme une fonction linéaire de l'ordre radial n. Pour tenir compte de variations plus locales tout en imposant un *a priori* de continuité, nous pouvons par exemple construire une fonction polynomiale d'ordre supérieur [2].

Dans le Chap. 4, chaque fréquence a été considérée comme indépendante des autres, ce qui a justifié l'utilisation des MCMC pour tenir compte des multiples minima suscités par l'accroissement du nombre de degrés de liberté. Ici, nous souhaitons que le calcul soit rapide et il est préférable d'utiliser du MAP. Cependant, nous cherchons aussi à ce que la solution trouvée soit unique. Pour assurer une meilleure stabilité, nous sommes donc obligés de réduire le nombre de degrés de liberté en restreignant le nombre de paramètres ou encore en appliquant des *a priori* forts. Ceci est d'autant plus vrai que les séries temporelles avec lesquelles nous comptons travailler peuvent être relativement courtes [3] et donc pauvres en information.

2. Une alternative, probablement plus fiable mais plus complexe à mettre en place, consisterait à utiliser des splines.

3. La plupart des étoiles Kepler ont été observées durant seulement 30 jours. Les observations se poursuivent actuellement pour les meilleures cibles sur des durées supérieures.

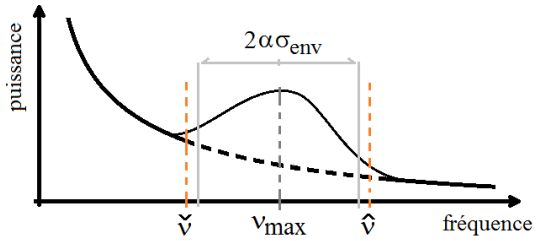


FIGURE 5.1 – Figure illustrant la manière dont est défini la plage sur lesquels les modes sont supposés significatifs.

On va imposer aux fréquences de suivre exactement un polynôme d'ordre 2 pour chaque degré l. La forme la plus évidente d'un tel polynôme est $\nu(n,l) = a_l n^2 + b_l n + c_l$. Cependant des tests ont révélé que l'ajustement ne fonctionne pas correctement avec cette expression : le résultat est systématiquement "figé" : le résultat de la minimisation est très fortement lié aux paramètres initiaux fournis. En fait, les différents termes de la somme sont très fortement corrélés entre eux et une petite variation du paramètre a_l peut être facilement compensée par des variations conjointes de b_l et c_l. Ceci est d'autant plus vrai que les données sont bruitées et que le coefficient de courbure moyen du motif de fréquences est faible.

Une solution consiste à modifier la paramétrisation. Supposons qu'il existe des corrélations entre a_l, b_l et c_l. Les différentes paires de corrélation sont, (a_l, b_l), (a_l, c_l), (b_l, c_l). Quel sera alors le terme de corrélation dominant et de quelle nature sera cette corrélation ? *A priori* nous ne pouvons répondre à ces questions. Cependant, en observant la relation précédente, on remarque que celle ci doit être probablement la plus forte entre les termes de degré les plus éloignés, a_l et c_l car une perturbation δa_l sur le terme a_l induit une variation $\delta\nu = \delta a_l n^2$ tandis qu'une perturbation δc_l sur le terme c_l induit une variation indépendante de n, $\delta\nu = \delta c_l$. De manière similaire, une perturbation δb_l sur le terme b_l induit une variation $\delta\nu = \delta b_l n$. Nous supposons donc que le terme dominant de corrélation est entre a_l et c_l.

Nous ne connaissons rien quant à la nature réelle de cette corrélation, qui très certainement varie d'une étoile à l'autre. On va la supposer linéaire : $c_l(a_l) = \alpha_l a_l + \gamma_l$. On néglige ainsi l'influence des corrélations (a_l, b_l) et (b_l, c_l) sur la capacité de l'algorithme d'ajustement à converger (ce qui ne signifie aucunement qu'elle ne sont pas significatives). La nouvelle paramétrisation du polynôme sera :

$$\nu(n,l) = a_l n^2 + b_l(a_l)n + c_l(a_l) = a_l n^2 + b_l(a_l)n + \alpha_l a_l + \gamma_l \tag{5.7}$$

On se retrouve avec 4 paramètres au lieu de 3. Mais on remarque aussi que pour $c_l(a_l = 0) = \gamma_l$. Cela signifie qu'un ajustement d'un polynôme du premier ordre nous permettrait de déterminer le terme γ_l. On se ramènerait alors au nombre de variable normal pour un polynôme d'ordre 2. En se basant sur la relation asymptotique, on identifie : $b_l(a_l = 0) = \Delta\nu$ et $c_l(a_l = 0) = \gamma_l = (\epsilon + l/2)\Delta\nu - l(l+1)D_0$.

Pour conclure, avec cette paramétrisation, nous sommes obligés de procéder en deux temps. Premièrement, on doit effectuer un ajustement polynômial d'ordre 1 dont les coefficients s'identifient à la loi asymptotique. On détermine ainsi γ_l. Deuxièmement, on fixe γ_l à la valeur fournie par le premier ajustement et on effectue un ajustement polynômial d'ordre 2 utilisant l'Eq.5.7.

Cette re-paramétrisation est une condition nécessaire pour que l'algorithme converge efficacement vers une solution stable et très souvent unique, en dépit des conditions difficiles de rapport

signal à bruit.

Pour le calcul des erreurs sur les fréquences individuelles, on doit calculer la matrice de covariance par inversion de hessienne (cf Chap.2.3). En supposant que seul les termes de fréquences sont corrélés, l'erreur s'écrira alors (loi de propagation des erreurs) :

$$\begin{aligned}\delta\nu(n,l)^2 = & \left(\frac{\partial\nu(n,l)}{\partial a_l}\right)^2 \delta a_l^2 + \left(\frac{\partial\nu(n,l)}{\partial b_l}\right)^2 \delta b_l^2 + \left(\frac{\partial\nu(n,l)}{\partial \alpha_l}\right)^2 \delta\alpha_l^2 + \\ & 2\frac{\partial\nu(n,l)}{\partial a_l}\frac{\partial\nu(n,l)}{\partial b_l} covar(a_l,b_l) + 2\frac{\partial\nu(n,l)}{\partial a_l}\frac{\partial\nu(n,l)}{\partial \alpha_l} covar(a_l,\alpha_l) + \\ & 2\frac{\partial\nu(n,l)}{\partial b_l}\frac{\partial\nu(n,l)}{\partial \alpha_l} covar(b_l,\alpha_l). \end{aligned} \tag{5.8}$$

A priori pour les fréquences

Maintenant que le choix du modèle pour les fréquences est effectué, nous devons définir l'ensemble des *a priori* que nous allons appliquer. Le choix de ceux-ci va reposer essentiellement sur la validité de la loi asymptotique (qui est supposée être une représentation assez bonne de la distribution des fréquences). Cette relation ne décrit correctement que les modes de pression dans les étoiles de type solaire dans la séquence principale car les étoiles plus évoluées ont des modes mixtes. Le champ d'application de notre méthode est donc restreint à cette catégorie d'étoile.

$\Delta\nu$ n'est très souvent connue qu'en moyenne et nous ne disposons pas d'évaluation au cas par cas de l'erreur la plupart du temps. Nous savons par contre que cette erreur est de l'ordre de 0.5% à 1% de $\Delta\nu$. On choisit donc d'utiliser un à priori gaussien centré sur l'estimation initial de la grande séparation dont :

- l'écart type est pris égal à 1% de $\Delta\nu$ si l'erreur est inconnue,
- l'écart type est pris égal à l'erreur à 1σ sinon.

Pour ϵ nous ne disposons que d'une estimation (ϵ^*). Cependant, nous pouvons tout de même établir des contraintes sur ce paramètre dont voici la liste :

- Premièrement, la loi asymptotique nous informe que dans tous les cas, $1 < \epsilon < 2$.
- Deuxièmement, ϵ contrôlant l'identification, les deux identifications doivent être clairement distinguées : on ne tolère pas l'inversion des modes pairs et impairs lors de l'ajustement.
- Troisièmement, on va exiger que les modes de degré l de même parité ne se croisent que difficilement.

La première contrainte est explicite. Qu'impose la seconde condition ?

Supposons que les grandeurs connues ne sont pas trop éloignées de la solution correcte. Introduisons alors les perturbations δ et $\delta\epsilon$ par rapport à $\Delta\nu^*$ et ϵ^*. Ainsi les valeurs "vraies" de ces paramètres peuvent s'écrire $\Delta\nu = \Delta\nu^* + \delta$ et $\epsilon = \epsilon^* + \delta\epsilon$.

Réécrivons alors la loi asymptotique au voisinage de la solution optimale, en tenant compte des Eq.5.2 et Eq.5.3,

$$\nu^* + \delta\nu = (n + \epsilon^* + \delta\epsilon + l/2)(\Delta\nu^* + \delta) - 0.02l(l+1)(\Delta\nu^* + \delta). \tag{5.9}$$

En ne préservant que les termes perturbatifs du premier ordre on a

$$\begin{aligned}\nu^* + \delta\nu = & (n + \epsilon^* + l/2)\Delta\nu^* - 0.02l(l+1)\Delta\nu^* + \\ & \delta\epsilon\Delta\nu^* + (n + \epsilon^* + l/2)\delta - 0.02l(l+1)\delta. \end{aligned} \tag{5.10}$$

Ce qui se simplifie en

$$\delta\nu = \delta\epsilon\Delta\nu^* + (n + \epsilon^* + l/2)\delta - 0.02l(l+1)\delta. \tag{5.11}$$

Nous avons dit que l'erreur sur la grande séparation était d'une fraction de pourcent. Prenons donc $\delta = 0.01\Delta\nu^*$. Ceci nous permet de réécrire la relation précédente uniquement en fonction de $\Delta\nu^*$

$$\delta\nu = \delta\epsilon\Delta\nu^* + 1.10^{-2}(n + \epsilon^* + l/2)\Delta\nu^* - 2.10^{-4}l(l+1)\Delta\nu^*. \tag{5.12}$$

Ne pas tolérer l'inversion des modes c'est imposer $\delta\nu < (\Delta\nu^* + \delta)\frac{1}{2}$, ce qui nous donne finalement une contrainte sur la variation $\delta\epsilon$ que l'on peut tolérer,

$$\delta\epsilon \approx \frac{1.01}{2} - 1.10^{-2}(n + \epsilon^* + l/2). \tag{5.13}$$

$\delta\epsilon$ peut donc être utilisé comme *a priori* pour empêcher une inversion des modes pairs et impairs. Nous avons choisi d'imposer un *a priori* uniforme sur ϵ tel que $\epsilon^* - \delta\epsilon < \epsilon < \epsilon^* + \delta\epsilon$.

On rappelle ici que ϵ est obtenu en moyenne, selon l'ordre radial n. Celui ci doit donc être recalculé à chaque itération de l'algorithme de minimisation.

Globalement, comme nous déterminons ϵ^* en utilisant le pic le plus haut, $n = n_{max} \approx \nu_{env}/\Delta\nu \approx 20 - 25$ pour les étoiles de type solaire (environ 22 dans le cas du soleil). Nous ne pouvons qu'estimer grossièrement ϵ puisqu'il est lié aux propriétés méconnues de la surface de chaque étoile. Dans le cas solaire, $\epsilon \approx 1.39$, ce qui nous donne $\delta\epsilon \approx 0.26$ pour le $l = 0$ et $l = 1$.

Dans la pratique, nous avons remarqué lors des tests, que $\delta\epsilon = 0.20$ est une valeur plus adaptée.

Comment tenir compte maintenant de la troisième condition ?

Les modes pairs et impairs de bas degré peuvent être définis relativement à la position du $l = 0$ et $l = 1$. Pour les modes de degré $l = 2$ (et éventuellement les $l = 3$ si ils étaient ajustés), on rappelle que la petite séparation entre $l = 0$ et $l = 2$ s'écrit : $\delta\nu_{02}(n) = \nu_{n+1,l=2} - \nu_{n,l=0} \approx 6D_0$.

Les modes définis par $\nu(n, 0)$ et $\nu(n, 2)$ sont donc en première approximation parallèles et tout écart à cette condition est le reflet de modification de grandeurs physique (densité ou vitesse du son par exemple) au sein de l'étoile. Se pose alors la question suivante : quel niveau de confiance pouvons nous donner à cette loi ? Ici, le choix subjectif que nous avons fait consiste à tolérer des variations de la quantité $\delta\nu_{02}(n)$ dans une gamme assez vaste au voisinage de la fréquence attendue. L'*a priori* utilisé ici est gaussien-uniforme-gaussien (cf. Fig.5.2). C'est le même *a priori* que celui utilisé lors de l'ajustement MCMC des étoiles CoRoT et Kepler dont l'identification est évidente (cf. Chap.4.1.2). Tel que défini, cet *a priori* a pour rôle essentiel de limiter la superposition des modes $l = 2$ et $l = 0$, mais sans l'exclure. L'écart type de la gaussienne, σ est calculé de telle sorte que chaque mode de degré $l = 0$ se situe exactement à 1σ, comme montré sur la Fig.5.2.

Les degrés $l > 3$ sont supposés d'intensité trop faible et ne sont pas pris en compte.

5.3.3 Contrainte de continuité des hauteurs et largeurs

Modélisation des hauteurs

Lorsque la résolution spectrale est médiocre, la détermination des hauteurs et des largeurs des modes en est profondément affectée. De plus, comme vu au Chap. 4.2.1, même dans des conditions favorables (haut rapport signal sur bruit, absence de superposition des modes pairs et impairs, splitting rotationnel résolu), les hauteurs et largeurs ne sont que grossièrement connues. Comme notre méthode vise justement à être appliquée plutôt dans des conditions difficiles, il est préférable de contraindre fortement hauteurs ou largeurs des modes. Mais comme ces deux paramètres sont anti-corrélés, il est inutile de les contraindre simultanément car cela risquerait d'introduire d'importants biais sur les amplitudes individuelles des modes.

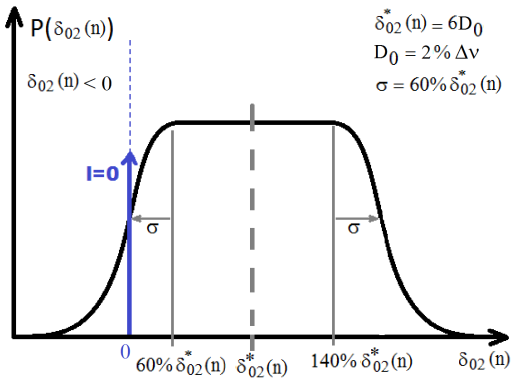


FIGURE 5.2 – *A priori* "gaussien-uniforme-gaussien" (GUG) appliqué à la petite séparation $\delta_{02}(n)$. δ^*_{02} correspond à la valeur attendu donnée par $6D_0$.

La paramétrisation des hauteurs étant plus simple que celle des largeurs, nous avons choisi d'appliquer des contraintes fortes sur celles ci. En observant l'évolution des hauteurs en fonction de l'ordre radial, on remarque que très souvent, celle ci suit une forme de cloche. C'est pourquoi on impose aux hauteurs de suivre exactement une loi gaussienne. Certaines étoiles connues (comme par exemple α *CenB*) ont un profil d'amplitude avec 2 maxima. Pour des étoiles de ce type, l'approximation gaussienne peut être trop forte. Une alternative (non testée) serait d'imposer à chacune des valeurs de hauteur (*i.e.* chacun des termes $H(n,l)$) de ne suivre qu'approximativement une gaussienne. Dans ce cas, les paramètres du profil de hauteur de l'*a priori* serait des hyper-paramètres, c'est à dire des *a priori* non fixés [4], dont les grandeurs sont ajustées durant la procédure de minimisation.

Dans le cas solaire, la hauteur des modes $l \neq 0$ est proportionnelle à la hauteur du $l = 0$. Nous supposerons que c'est aussi le cas des autres étoiles de type solaire et n'ajusterons qu'une seule fonction gaussienne (Eq. 5.14). Les hauteurs des modes $l \neq 0$ seront définies relativement à celle du $l = 0$, comme dans tous les exemples présentés jusque là. On nommera les paramètres hauteur maximum, largeur et fréquence centrale des modes respectivement, H_{modes}, σ_{modes}, ν_{modes} (à ne pas confondre avec les paramètres de l'enveloppe H_{env}, σ_{env}, ν_{env} lors de l'étape précédente). On a donc,

$$H(\nu, l=0) = H_{modes} e^{-0.5(\nu-\nu_{modes})^2/\sigma^2_{modes}} \quad (5.14)$$

$$H(\nu, l=1) = V_{l=1} H(\nu, l=0) \quad (5.15)$$

$$H(\nu, l=2) = V_{l=2} H(\nu, l=0). \quad (5.16)$$

A priori pour les hauteurs

Nous n'utilisons ici que des *a priori* uniformes pour contraindre les paramètres de la gaussienne décrivant les hauteurs du $l = 0$. Les bornes de ces fonctions sont définies en se basant sur les résultats de l'étape d'ajustement de l'enveloppe des modes. Le tableau 5.2 résume les valeurs caractéristiques utilisées.

4. Par exemple si on a un *a priori* gaussien $N(\alpha, \sigma)$ pour une variable x, cela revient à estimer α et σ en même temps que x.

TABLE 5.2 – Synthèse des *a priori* sur les hauteurs utilisés lors de l'ajustement polynomial (ordre 1 et 2). On rappelle que y correspond au spectre de puissance.

paramètre	*a priori*	Unité
H_{modes}	$U(0,\, max(y))$	$ppm^2/\mu Hz$
ν_{modes}	$U((n_- - 1)\Delta\nu^*,\, (n_+ + 1)\Delta\nu^*)$	μHz
σ_{modes}	$U(0.25\sigma_{env},\, 1.5\sigma_{env})$	μHz

Nous aurions pu utiliser des *a priori* gaussiens puisque nous disposons des barres d'erreurs sur les paramètres de l'enveloppe des modes et de la relation d'échelle donnée par l'Eq.1.66 reliant hauteur des modes (contrôlé par H_{modes}) et hauteur de l'enveloppe des modes (contrôlé par H_{env}). Mais cela est trop contraignant : des tests utilisant l'Eq.1.66 comme *a priori* ont montré que la qualité des résultats était parfois dégradée et que pour certaines étoiles la convergence de l'ajustement n'était plus assurée : cette relation est probablement trop grossière. Par contre, elle est suffisante pour être utilisée pour définir la hauteur initiale des modes avant ajustement.

Les *a priori* pour les hauteurs relatives $V_{l=1}$ et $V_{l=2}$ sont définis sur la base de trois critères :

- la hauteur du $l = 1$ est supérieure à celle du $l = 0$ mais n'excède pas deux fois celle ci : $H(\nu, l = 0) < H(\nu, l = 1) < 2H(\nu, l = 0)$.
- la hauteur du $l = 2$ est positive et inférieure à celle du $l = 0$: $0 < H(\nu, l = 2) < H(\nu, l = 0)$.
- le maximum de probabilité des hauteurs relatives $V_{l=1}$ et $V_{l=2}$ correspond à celui du cas solaire.

Pour satisfaire à ces exigences, nous appliquons un *a priori* gaussien-tronqué (cf. Annexe.A) sur les hauteurs relatives $V_{l=1}$ et $V_{l=2}$. Nous ne bénéficions que de peu d'études sur de possibles variations de la hauteur relative des modes. les hauteurs relatives solaires sont obtenues en tenant compte du facteur de visibilité géométrique des modes après intégration sur tout le disque stellaire, et de l'effet centre bord. Ce dernier coefficient peut varier en fonction de l'atmosphère stellaire. On s'attend à des variations de l'ordre de 5 à 10 %. Cependant, une étoile CoRoT semble avoir des hauteurs relatives inhabituelles : HD49385. Pour cette étoile plutôt évoluée il semblerait que le degré $l = 3$ soit parfois visible. Selon Deheuvels et al. (2010) nous avons,

$$\begin{aligned} V_{l=1} &= H_{l=1}/H_{l=0} = 1.54 \pm 0.16 \\ V_{l=2} &= H_{l=2}/H_{l=0} = 0.76 \pm 0.12 \\ V_{l=3} &= H_{l=3}/H_{l=0} = 0.20 \pm 0.05 \end{aligned}$$

Le rapport signal à bruit de HD49385 est élevé et les modes $l = 0$ et $l = 2$ sont clairement dissociés : l'identification est évidente. Par contre les modes $l = 2$ et $l = 3$ ont des hauteurs relatives très supérieure à celle attendues. L'atmosphère doit probablement jouer un rôle bien plus significatif que prévu pour certaines étoiles. Nos tests présentés dans les chapitres suivants ont portés sur le Soleil et sur un ensemble d'étoiles simulées issues d'un catalogue mis à disposition par la communauté travaillant sur Kepler [5].

Pour le Soleil (cf. un peu plus bas), notre connaissance sur la hauteurs relatives des modes est précise. Nous utilisons dans ce cas un *a priori* gaussien dont les grandeurs caractéristiques sont résumées Tableau 5.4.

5. Ce sont des étoiles simulées spécifiquement pour effectuer des exercices de performance d'algorithmes d'extraction de paramètres stellaire. Ce catalogue est représentatif de ce que Kepler est capable de produire en terme de courbe de lumière.

TABLE 5.3 – Synthèse des *a priori* sur les hauteurs relatives utilisées lors de l'ajustement polynomial (ordre 1 et 2) dans le cas stellaire. L'*a priori* "gaussien-tronqué" est représenté en Annexe A.

paramètre	*a priori*
$V_{l=1}$	$GT(1.50,\ 0.5,\ 1,\ 2)$
$V_{l=2}$	$GT(0.53,\ 0.2,\ 0,\ 1)$

TABLE 5.4 – Synthèse des *a priori* sur les hauteurs relatives utilisées lors de l'ajustement polynomial (ordre 1 et 2) dans le cas solaire. Deux écarts type ont été testés pour évaluer l'influence de l'*a priori*. L'*a priori* gaussien est représenté Annexe A.

paramètre	*a priori*
$V_{l=1}$	$N(1.50,\ 0.10)$ ou $N(1.50,\ 0.20)$
$V_{l=2}$	$N(0.53,\ 0.05)$ ou $N(0.53,\ 0.10)$

Pour les autres étoiles, les ajustements que nous avons menés se servent de HD49385 comme référence pour définir le degré de tolérance de la fonction gaussienne-tronqué (GT) (*i.e.* paramètre σ de la Table 5.3).

Cette *a priori* est non normalisé. Donc il ne peut être utilisé pour calculer une probabilité de modèle par échantillonnage MCMC (Chapitre 2.2). A partir du moment où il est inchangé d'une identification à l'autre (constante de normalisation identique), il ne pose par contre aucun problème avec une approche de type MAP.

5.3.4 Le bruit lors de l'ajustement des modes

Tout comme dans le chapitre 5.2, où l'on cherchait à ajuster l'enveloppe des modes, le bruit est modélisé par 2 profils de Harvey et un bruit blanc (Eq.5.1a).

Les *a priori* sur le bruit sont définis en se servant des informations collectées lors de l'ajustement de l'enveloppe des modes (valeur du maximum de densité de probabilité pour chaque paramètre et erreur estimée au voisinage du maximum).

Les paramètres associés au profil de Harvey dominant à plus basse fréquence sont fixés. Ceci est équivalent à utiliser des distributions de Dirac comme *a priori*.

Le second profil de Harvey et le bruit blanc sont contraints par des *a priori* gaussiens. Dans les étoiles de type solaire de la séquence principale, c'est le profil qui contribue le plus au bruit dans la région où les modes sont excités. Les maxima de densité de probabilité trouvés chapitre 5.2 définissent la valeur la plus probable et l'écart type de l'*a priori* est défini comme 3 fois l'erreur estimée.

5.3.5 Remarque sur l'angle d'inclinaison et le splitting rotationnel

L'angle d'inclinaison a une incidence importante sur la hauteur des composantes $\pm m$ lorsque le splitting rotationnel est résolu. Dans le cas solaire, la rotation étant d'environ 28 jours, deux composantes consécutives de l'ordre azimutal ($m = 0$ et $m = 1$ par exemple) sont séparées d'environ 0.4 μHz. Il est donc inutile de tenir compte du splitting pour des observations plus courtes que 28 jours. En pratique, il est impensable d'obtenir une information fiable à une

résolution spectrale de 0.4 μHz. La résolution nécessaire est au moins 2 ou 3 fois plus élevée car la largeur des modes induit leur superposition.

Comparé au cas stellaire dans sa généralité, le Soleil est un rotateur plutôt lent et il est commun d'avoir des étoiles tournant en quelques jours comme pour HD49933 et HD181420. Cependant, nous avons vu qu'à faible rapport signal à bruit, ajuster le splitting est problématique car il peut exister de multiples maxima sous l'influence du bruit ou de propriétés stellaires.

Enfin, les approches basées uniquement sur le maximum du critère statistique (MLE ou MAP) rendent encore plus difficile la détermination du couple angle/splitting. Il faut disposer de conditions initiales (*initial guesses*) relativement précises, ce qui n'est pas aisé. Par exemple, pour les étoiles observées par Kepler, une estimation du splitting n'est la plupart du temps pas disponible car les séries temporelles sont filtrées à basse fréquence pour éviter une découverte inopportune d'exoplanète. Dans ce cas, aucune information *a priori* ne nous permet de décorréler angle d'inclinaison et rotation (on ne connait alors qu'au mieux le couple $vsin\ i$ de surface).

En conséquence, nous avons décidé pour les cas présentés ici de ne pas tenir compte de l'angle d'inclinaison stellaire et du splitting rotationnel (tous deux fixés à 0). Ceci revient ne pas tenir compte des multiplets m lors de l'ajustement. Si à l'avenir il s'avérait possible d'obtenir de manière systématique et fiable un indicateur sur l'une ou l'autre de ces grandeurs, alors il serait facile d'apporter les modifications nécessaires à notre code.

Nous n'attendons pas un impact significatif sous forme de biais de l'absence de ces paramètres sur les fréquences car, à l'exception des rotateurs très rapides, les composantes "splittés" sont disposées peu ou prou de manière symétrique autour du $m = 0$. Par contre, les hauteurs et les largeurs des modes de degré $l > 0$ seront affectées.

Les conséquences de l'approximation brutale effectuée sur le couple angle et splitting seront entrevues dans les chapitres suivants, grâce aux simulations menées.

5.3.6 Résumé de la méthode

Nous avons construit différents modules permettant d'effectuer les taches précédemment décrites. Nous résumons ici la logique globale et le processus décisionnel que nous avons choisi pour extraire automatiquement les modes de bas degré $l = 0$, $l = 1$, $l = 2$ d'une part dans le cas stellaire et d'autre part dans le cas solaire.

On procède tout d'abord à un "nettoyage" sommaire des séries temporelles. Ainsi, les sauts d'intensité dus essentiellement aux impacts de protons sur le CCD et les dérives instrumentales sont enlevés.

Ensuite, la fonction densité spectrale de puissance est calculée et les paramètres initiaux sont calculés en se servant des relations d'échelle décrites Chapitre 1.2.8. Pour le bruit, nous nous servons directement du spectre pour calculer des paramètres initiaux grossiers. Tous les paramètres sont alors affinés en utilisant un algorithme d'exploration de type MCMC.

L'ajustement proprement dit peut alors être effectué. Nous cherchons tout d'abord à valider/invalider la présence de modes stellaires significatifs au regard du bruit. Pour cela, nous comparons deux modèles en se servant d'un algorithme de minimisation de type maximum *a posteriori*. Le premier modèle comporte à la fois des modes dont l'enveloppe est décrite par une fonction gaussienne, et une description du bruit stellaire et instrumentale. Le second modèle ne comporte que du bruit.

Nous acceptons toutes les étoiles pour lesquelles la probabilité de présence des modes est supérieure à 50%. Les étoiles sélectionnées sont transmises à une autre routine chargée de l'extraction des paramètres individuels des modes. Le modèle décrivant les dits modes est pensé pour assurer une mesure assez précise des fréquences des modes en accord avec les connaissances

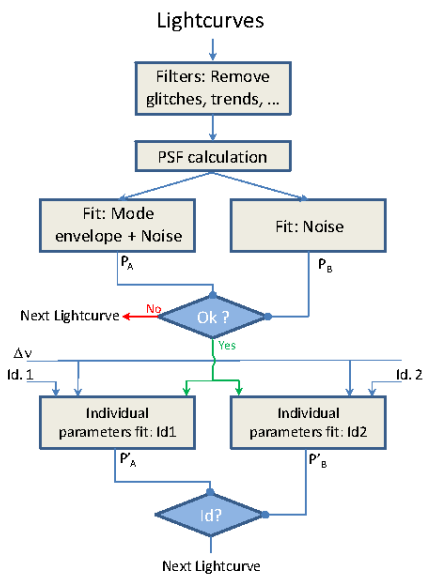


FIGURE 5.3 – Représentation schématique de l'algorithme utilisé lors de l'ajustement des modes stellaires. Les losanges correspondent aux étapes pour lesquels un test logique (présence de modes ?, identification ?) est effectué.

a priori que nous avons des propriétés de pulsations des étoiles de type solaire dans la séquence principale. La grande séparation est pour le moment supposée estimée par un algorithme externe (table fournie en entrée de l'algorithme).

Deux cas se présentent alors. Si nous ajustons des données solaires, une seule identification est ajustée puisque celle ci est parfaitement connue pour cet astre, et les *a priori* utilisés sont très contraignants pour les hauteurs relatives des modes.

Si nous travaillons avec des données stellaires, les deux identifications des degrés sont envisagées. A l'issue de l'ajustement, les deux identifications sont confrontées *via* un test de vraisemblance (calcul du rapport concurrentiel au sens défini au Chap. 2.2.3) et l'identification la plus probable ainsi obtenu est retenue.

Finalement, les erreurs et la matrice de covariance sont estimées en se basant sur la Hessienne.

L'ensemble du processus décisionnel dans le cas de l'analyse stellaire est schématisé Fig.5.3.

5.4 Test sur les données solaires : mise en évidence du cycle solaire

Pour tester les capacités de notre méthode, nous avons dans un premier temps analysé les données de l'instrument VIRGO (Frohlich et al. 1997). Ce sont des séries temporelles de l'intensité solaire (photométrie) acquises par le satellite SoHo, toujours en fonctionnement actuellement. Nous disposons ainsi de plus de 15 ans de données quasi-continues.

Pour disposer de séries temporelles de précision comparable à la plupart de celles acquises lors du début de la mission Kepler, nous avons choisi de découper les 15 ans en sous-séries d'une durée de 30 jours. Par la suite, chaque sous-série est analysée comme décrit dans les parties précédentes.

Suite à une perte du contrôle du satellite durant 2 mois, en début de mission, nous avons un trou dans les données. De plus, certaines séries sont ponctuées d'interruptions plus ou moins courtes. La résolution spectrale est donc au mieux de 0.386 μHz.

Les spectres de puissance des sous-séries temporelles ont un très bon rapport hauteur sur bruit. Ce rapport est de l'ordre de 20 à 40 au voisinage de la fréquence centrale $\nu_{env} \approx \nu_{modes}$.

Comme paramètre d'entrée, nous connaissons évidement déjà ν_{env} et donc A_{max}. La grande séparation moyenne est aussi connue. Nous avons retenu $\Delta\nu = 135\mu$Hz.

A l'issue de l'ajustement, nous avons regardé si toutes les séries temporelles avaient été considérées comme ayant des modes : ce fut le cas. Cela signifie que toutes les courbes de lumière valides[6] passent le premier test logique du schéma de la Fig.5.3.

On procède alors à l'ajustement des paramètres des modes individuels sur la base des contraintes exposées dans les chapitres précédents. Les Fig.5.4 montrent un exemple de spectre de puissance et de diagramme-échelle obtenu pour une sous-série de 30 jours de durée. La répartition en fréquence des modes solaires est remarquablement bien reproduite par un polynôme d'ordre 2, autour du maximum d'amplitude. Nous avons ensuite regardé la dépendance temporelle des paramètres ajustés pour déterminer si la méthode était suffisamment sensible pour mettre en évidence le cycle d'activité du soleil.

Sur la Fig.5.5 on a représenté l'évolution temporelle de la fréquence du mode $l = 0$ dans la région de plus haut rapport hauteur sur bruit ($n = 21$ ici). On voit clairement la modulation périodique due à l'activité du Soleil. Ce type de modulation est bien connue depuis qu'elle a été mise en évidence par Woodard & Noyes (1985). Par contre, il est remarquable de pouvoir extraire des variations de l'ordre de la résolution spectrale sur un mode individuel alors que la description utilisée pour les fréquences est globale. Bien entendu, la dispersion est relativement importante mais la tendance générale des variations est clairement visible sur une large gamme de modes (plage de fréquence $\nu \in [2500, 3900]$ μHz soit une dizaines d'ordres radiaux) et pour les 3 degrés ajustés. L'erreur typique calculé *via* l'Eq.5.8 est de l'ordre de $0.1 - 0.2$ μHz.

Des travaux antérieurs tels que ceux de Chaplin et al. (2000) ou encore Komm et al. (2000) ont aussi mis en évidence des variations des valeurs caractéristiques de hauteurs et largeurs des modes durant le cycle. Nos ajustements spectraux montrent aussi cette tendance pour les largeurs. Cependant, nous sommes limités par la résolution spectrale et par le rapport hauteur sur bruit. Ainsi cet effet n'est visible que lorsque la largeur des modes est supérieure à la résolution spectrale (cet limitation concerne donc les basses fréquences) et lorsque lorsque le rapport hauteur sur bruit est suffisant, ce qui exclut les modes d'ordre radial supérieur à 26. Sur la Fig.5.6 sont montrés les variations cycliques de largeur pour l'ordre radial $n = 21$.

Bien que le profil des hauteurs suit exactement une loi gaussienne, l'effet de l'activité a pu être mis en évidence. La Fig.5.7 montre les variations du paramètre H_{modes} (*i.e.* la hauteur maximale de la gaussienne) au cours du cycle.

On précisera aussi que l'amplitude des modes définie comme $\sqrt{\pi H \Gamma}$ décroit et croit de concert avec l'activité comme déjà signalé dans Chaplin et al. (2000), bien que les variations de hauteur et de largeur soient en opposition de phase (Fig.5.8). On confirme aussi les travaux de Chaplin et al. (2000) sur le taux d'excitation des modes : celui-ci défini comme $\sqrt{\pi H \Gamma^2}$, est constant tout le long du cycle (aux incertitudes près).

Cette méthode a permis de mettre en évidence une conséquence jusqu'alors ignorée du cycle d'activité sur les modes p. La hauteur relative du $l = 2$ ($V_{l=2}$) diminue lorsque l'activité croit. Ces variations très faibles, sont de l'ordre de 5% mais la tendance globale est clairement visible en Fig.5.9. Il est vraisemblable que l'amortissement de ces modes soit plus sensible que les autres

6. A cause de l'interruption de 2 mois il n'y aucun point valide sur deux des sous-séries. Dans ce cas aucune analyse n'est évidement possible.

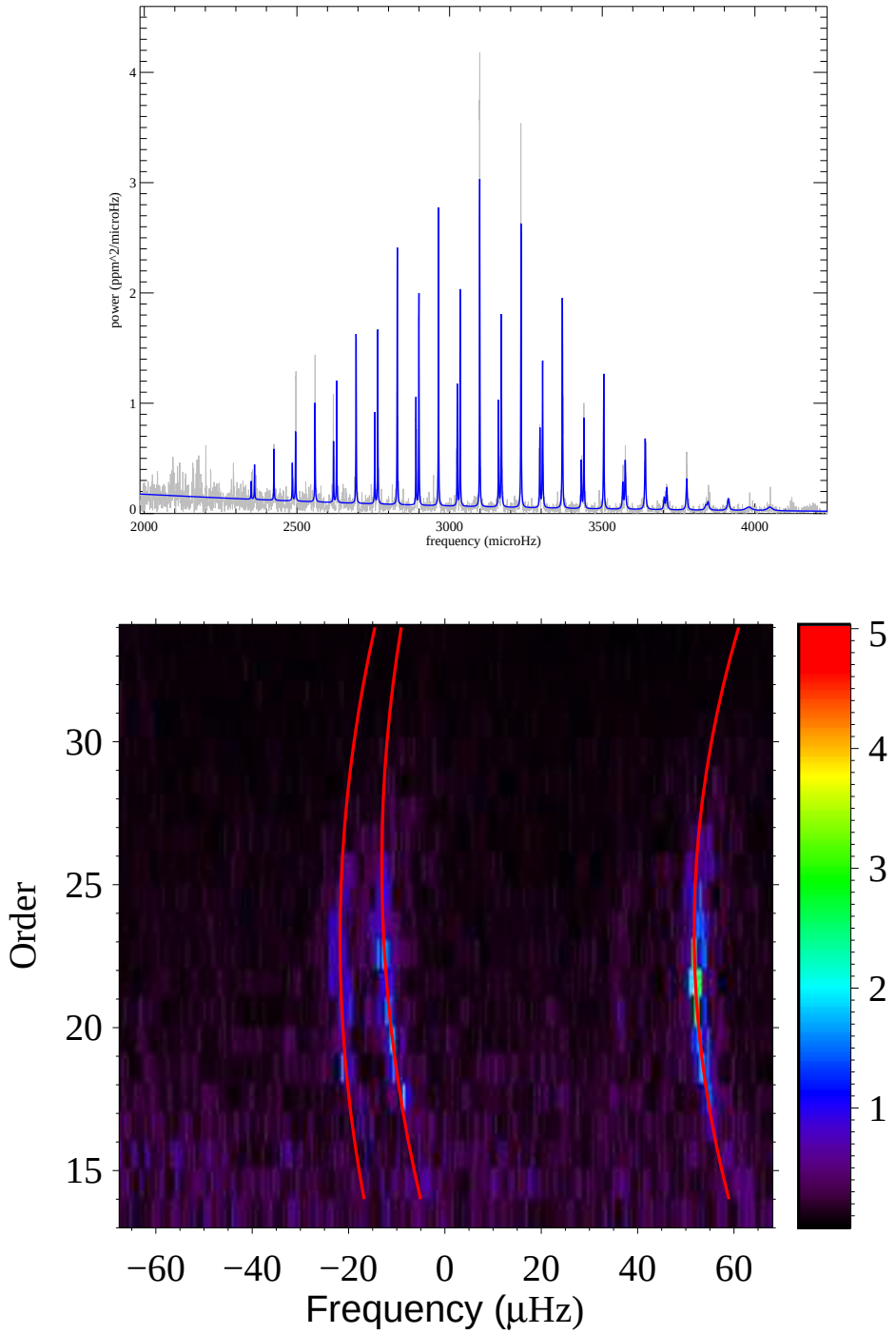


FIGURE 5.4 – Exemple de spectre solaire et de diagramme échelle obtenus avec une sous-série de 30 jours de données de l'instrument VIRGO. Sur le spectre le meilleur ajustement obtenu est représenté par une ligne bleue. Le rapport hauteur sur bruit est très élevé et la distribution en fréquence des modes très bien reproduite par un polynôme d'ordre 2 (lignes rouges sur le diagramme échelle). Sur le diagramme échelle on distingue à gauche le couple $l = 0$ et $l = 2$ et à droite le $l = 1$.

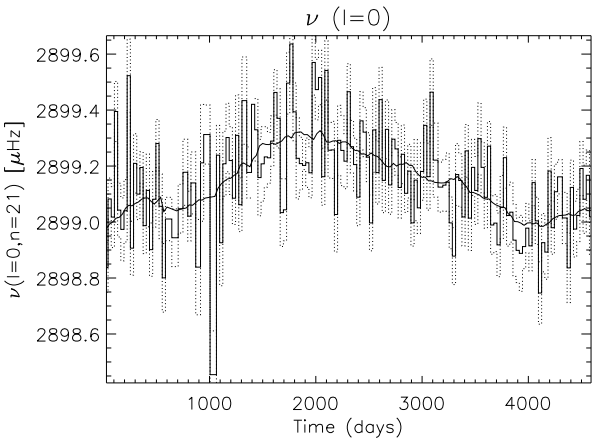


FIGURE 5.5 – Evolution temporelle de la fréquence du mode $n = 21, l = 0$ (plus haut rapport signal à bruit). Les barres d'erreurs à 1σ sont représentées par les courbes pointillées. Une variation cyclique est clairement visible. La méthode est suffisamment sensible pour détecter des variations fréquentielles de l'ordre de 0.2-0.4 μHz au cours du cycle. Une courbe de tendance (moyenne glissante) a été superposée au observations. La date d'observation débute le 1er Mars 1996.

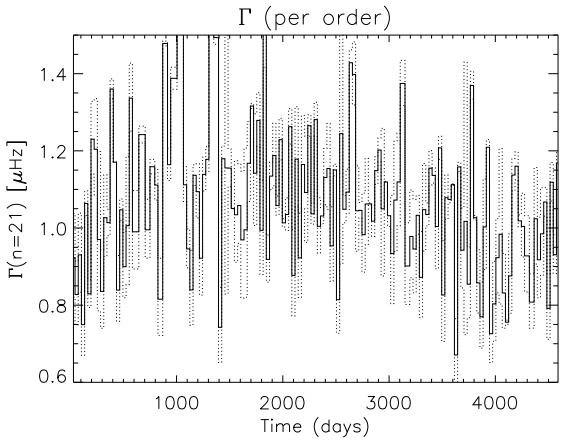


FIGURE 5.6 – Evolution temporelle de la largeur pour l'ordre $n = 21$. La dépendance vis à vis de l'activité solaire est visible. Les lignes pointillées correspondent à l'intervalle à 1σ.

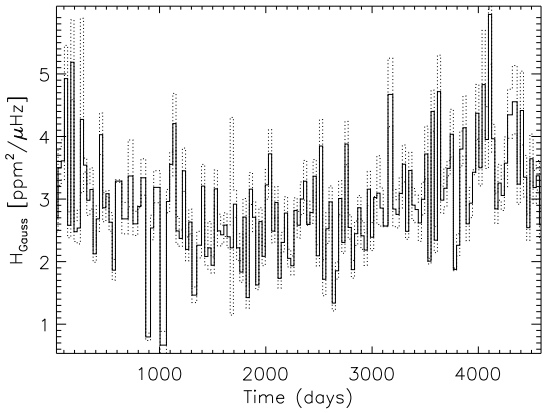


FIGURE 5.7 – Evolution temporelle de la hauteur de l'enveloppe des modes. Une dépendance au travers du cycle d'activité solaire est visible. Les lignes pointillées correspondent à l'intervalle à 1σ.

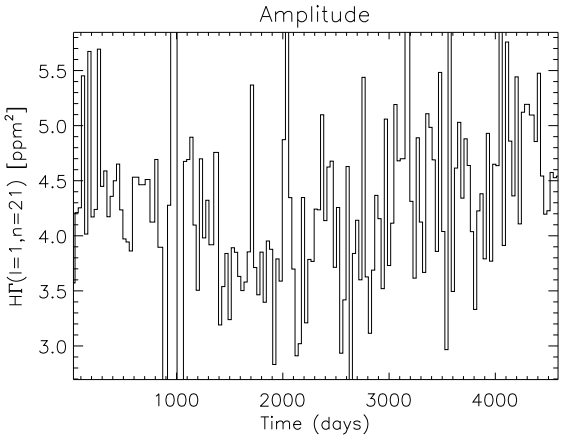


FIGURE 5.8 – Evolution temporelle de l'amplitude de l'ordre $n = 21$. L'amplitude fluctue au gré de l'activité solaire.

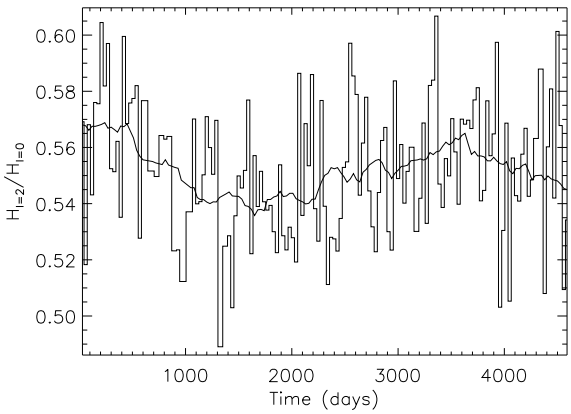


FIGURE 5.9 – Evolution temporelle de la hauteur relative du degré $l = 2$. Une dépendance au travers du cycle d'activité solaire est visible. Une courbe de tendance est superposée aux résultats.

aux effets du champ magnétique des taches solaires, plus nombreuses près de l'équateur, où les modes $l = 2$ se concentrent. Par opposition, les $l = 0$ et $l = 1$ se répartissent selon toutes les latitudes.

Au final, dans des conditions idéales de mesures, notamment de rapport hauteur sur bruit, la méthode est capable d'extraire les variations infimes des paramètres individuels des modes p, malgré une description sommaire de ceux ci.

5.5 Performance de l'algorithme sur un échantillon d'étoiles simulées

Fort des résultats encourageants obtenus avec les données solaires, nous avons cherché à estimer la qualité de l'ajustement sur un ensemble de courbes de lumières simulées. Dans le cadre de la mission Kepler, des courbes de lumière mimant le comportement (bruit de granulation, activité, tout type de modes de pulsations, effet rotationnels, ...) d'une large gamme d'étoiles ont été produites. Ces données, souvent nommées AsteroFLAG (nom d'un groupe internationnal dont l'objectif est la caractérisation des spectres des étoiles de type solaire) sont décrites dans Chaplin et al. (2008) et constituent un bon moyen de test de notre méthodologie. Nous avons utilisé 177 étoiles simulées dont les magnitudes s'étalent de 7 à 10.8. Chacune des séries temporelles a une durée de 30 jours.

Contrairement au cas solaire, nous ne disposons pas d'approximation *a priori* de la grande séparation. Idéalement, celle-ci devrait être fournie par un algorithme tiers que nous n'avons pas développé. Dans le cas des données simulées, nous disposons des fréquences vraies de chacune des 177 étoiles simulées. Nous avons donc calculé la grande séparation moyenne sur la base de ces fréquences, puis ajouté une perturbation aléatoire gaussienne, dont l'écart type est de 0.4% de $\Delta\nu$. Notons que lorsque nous utilisons un écart type de 1% $\Delta\nu$, des problèmes de convergences apparaissent pour certaines étoiles. Il est donc crucial d'avoir une très bonne estimation de la grande séparation moyenne pour pouvoir ajuster les modes.

Sur les 177 étoiles simulées, 79 (soit 44.6%) ont un signal jugés significatif au regard du bruit (*i.e.* 79 étoiles passent le premier test logique). La detection des modes p dépend très fortement

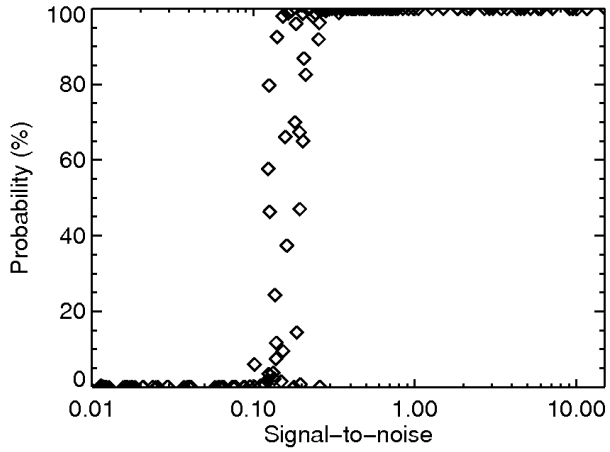


FIGURE 5.10 – Probabilité de présence des modes en fonction du rapport signal à bruit dans le cas des données simulées AsteroFLAG. Un clair seuil est visible autour de $H_{env}/B_{env} \approx 0.15$, B_{env} étant le niveau de bruit à la position du maximum de l'enveloppe ν_{env}. En dessous de ce seuil, les étoiles sont rejetées : elles ne feront pas l'objet d'une analyse plus approfondie.

du rapport signal à bruit comme illustré par la Fig.5.10. Le seuil de réjection visible autour de $H_{env}/B_{env} \approx 0.15$, B_{env} étant le niveau de bruit à la position du maximum de l'enveloppe ν_{env} est susceptible de changer. En effet, en modifiant les *a priori* on peut controler la position et l'étendue de la transition entre étoiles rejetées et acceptées. On fera remarquer qu'avec les *a priori* actuels, 3% des étoiles sont considérées comme n'ayant pas d'excès d'énergie significatif en dépit d'un léger excès de puissance visible à l'oeil. Elles correspondent donc à des cas de "faux négatif". On met aussi en evidence une dépendance directe entre magnitude des étoiles et taux de réjection comme nous pouvions nous y attendre (Fig.5.11).

Sur la Fig.5.12 on a représenté l'erreur relative effectuée sur la détermination de la fréquence centrale de l'enveloppe des modes. Celle ci est définie par la quantité $(\nu_{env}^{out} - \nu_{env}^{in})/\nu_{env}^{in}$ où ν_{env}^{in} et ν_{env}^{out} représentent respectivement la position centrale véritable des modes p et la valeur mesurée de cette même quantité. La valeur moyenne de la distribution des erreurs est de 0.3% pour un écart type de 3.4% si bien que si biais il y a, ce dernier est très inférieur à l'incertitude. Un éventuel biais pourrait résulter de l'approximation par une enveloppe gaussienne des modes.

Les deux identifications possibles sont comparées grâce à un rapport concurrentiel comme déjà signalé dans les paragraphes précédents. Le seuil de réjection est fixé à 90%. Cela signifie que l'identification dominante est supposée certaine que si sa probabilité dépasse ce seuil. Au final, sur les 79 étoiles restantes à l'issue du premier test logique, nous constituons 4 catégories,

Catégorie 1. les étoiles dont l'identification est correctement déterminée par le test statistique. Dans ce cas la probabilité dépasse 90% pour l'une ou l'autre des identifications. 55 étoiles sont dans ce cas.

Catégorie 2. les étoiles pour lesquelles il n'a pas été possible d'obtenir une identification des modes de manière non ambiguë. Dans ce cas il n'a pas été possible d'atteindre le seuil fatidique de 90% pour l'une ou l'autre des identifications. 8 étoiles sont dans ce cas.

Catégorie 3. les étoiles pour lesquelles l'identification est incorrecte à l'issue du test statistique. C'est le cas de 7 étoiles.

Catégorie 4. les étoiles pour lesquelles le test statistique n'a pu être effectué. C'est le cas de 9 étoiles. Dans les présentes simulations la raison est que lors de la recherche du maximum

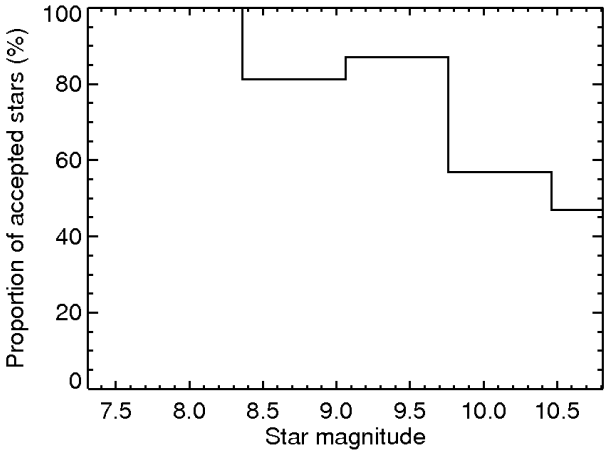


FIGURE 5.11 – Proportion d'étoiles acceptées à l'issue de la recherche d'excès de puissance significatif pour les données simulées AsteroFLAG. Le nombre d'étoiles acceptées est représenté en fonction de la magnitude. Elles ont été regroupées par tranche de 0.7 magnitude. Comme attendue, plus la magnitude est élevée, plus le nombre d'étoiles acceptées est faible.

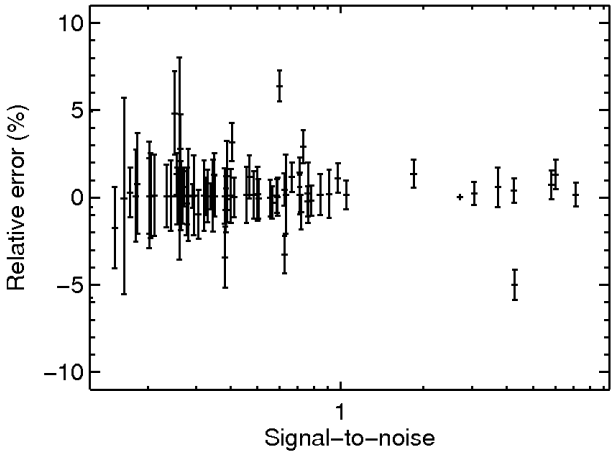


FIGURE 5.12 – Différence relative entre la fréquence centrale véritable des modes p et la valeur mesurée (quantité $(\nu_{env}^{out} - \nu_{env}^{in})/\nu_{env}^{in}$) en fonction du rapport signal à bruit obtenu après ajustement de l'enveloppe des modes par une gaussienne. La moyenne et l'écart type de la distribution sont respectivement 0.3% et 3.4%.

de probabilité, l'algorithme se trouve dans l'incapacité de converger (arrêt prématuré).

Toutes les étoiles de la dernière catégorie ont une fréquence centrale des modes inférieures à 900 μHz. Sur les Fig.5.13, on a représenté l'écart quadratique moyen (RMS) sur les résultats de l'estimation des fréquence, de l'amplitude et de la largeur. Evidemment, idéalement cet écart RMS doit valoir zéro. On voit clairement apparaitre un seuil à 900 μHz sur ces figures. A l'évidence, l'ajustement ne converge pas correctement, ou lorsqu'il converge, le résultat est entaché de grosses erreurs car l'hypothèse de continuité des fréquences est caduque (notamment pour le degré $l = 1$). Les étoiles les plus évoluées ont par exemple des modes mixtes et ne peuvent être correctement ajustées. Or celles ci ont une fréquence centrale relativement faible. L'accroissement du nombre de modes mixtes et la violation de la loi asymptotique qui en résulte sont aussi visibles dans les données Kepler, avec un seuil pour la fréquence centrale de l'enveloppe des modes similaire, autour de 900 μHz.

Sur les 70 étoiles pour lesquelles le calcul de probabilité a été possible (catégorie 1, 2 ou 3), environ 78% sont correctement ajustées par la loi asymptotique. Nous ne pouvons aboutir à une conclusion pour 12% d'entre elles et 10% sont incorrectement identifiées. Sur la Fig.5.15 on pourra apprécier visuellement comment se répartissent les étoiles des 4 catégories en fonction du rapport hauteur sur bruit maximum du $l = 0$, et de la fréquence centrale des modes.

La plupart des étoiles incorrectement identifiées ont un rapport maximum de hauteur sur bruit relativement faible, à l'exception d'un cas. L'examen de ce cas particulier nous permet d'identifier la cause de l'échec du test statistique : les modes de degrés $l = 2$ sont incorrectement ajustés. En effet, la procédure d'identification dépend très fortement de la capacité à ajuster correctement ce mode. Il est raisonnable de penser qu'en augmentant la contrainte sur la hauteur relative du $l = 2$ (quantité $V_{l=2}$), nous pourrions mieux discriminer entre les deux identifications des modes possibles.

Sur les Fig.5.16 et Fig.5.17 on présente un exemple d'ajustement obtenus lorsque les degrés des modes sont correctement identifiés. Même dans des conditions de rapport hauteur sur bruit particulièrement faible (*i.e.* < 4), il est possible d'obtenir une estimation raisonnablement bonne des fréquences, ce qui est justement le but recherché. L'erreur relative, défini comme $(\nu_{n,l}^{out} - \nu_{n,l}^{in})/\nu_{n,l}^{in}$ est par exemple inférieure à 0.05% pour le $l = 1$ pour les modes de rapport hauteur sur bruit supérieure à 10. En conséquence la précision est proche de la résolution spectrale. Pour des hauteurs sur bruit inférieures, l'erreur relative s'acroit, mais reste relativement bonne. Les modes de degré $l = 0$ et plus particulierement les modes $l = 2$ sont moins intenses que les $l = 1$ et par conséquent la précision obtenue est inférieure. On pourra apprécier visuellement la dispersion de l'erreur relative en fonction du rapport hauteur sur bruit sur la Fig.5.14.

Pour la hauteur et la largeur des modes, la précision obtenue est bien moindre, particulièrement à bas rapport hauteur sur bruit. L'erreur typique est de l'ordre de la dizaine à la centaine de pourcent dans le pire des cas. La dispersion est particulièrement grande pour $l > 0$, très probablement parce que les composantes rotationnelles des modes l ne sont pas ajustées. On rappellera aussi que le profil des hauteurs est contraint à suivre une fonction gaussienne, ce qui doit conduire inévitablement à des biais. Cette méthode ne permet donc pas d'obtenir de valeur précise de hauteur et de largeur, mais seulement un ordre de grandeur à cause de la faible résolution des spectres. Notons qu'avec une résolution supérieure, les résultats serait probablement meilleurs. Enfin, les amplitudes sont bien contraintes pour $l = 0$ et bien moins pour $l > 0$, encore une fois parce que les composantes rotationnelles sont négligées. On pourra apprécier visuellement les dispersions des distributions des erreurs relatives pour les hauteurs, largeurs et amplitudes sur les Fig.5.18, Fig.5.19 et Fig.5.20.

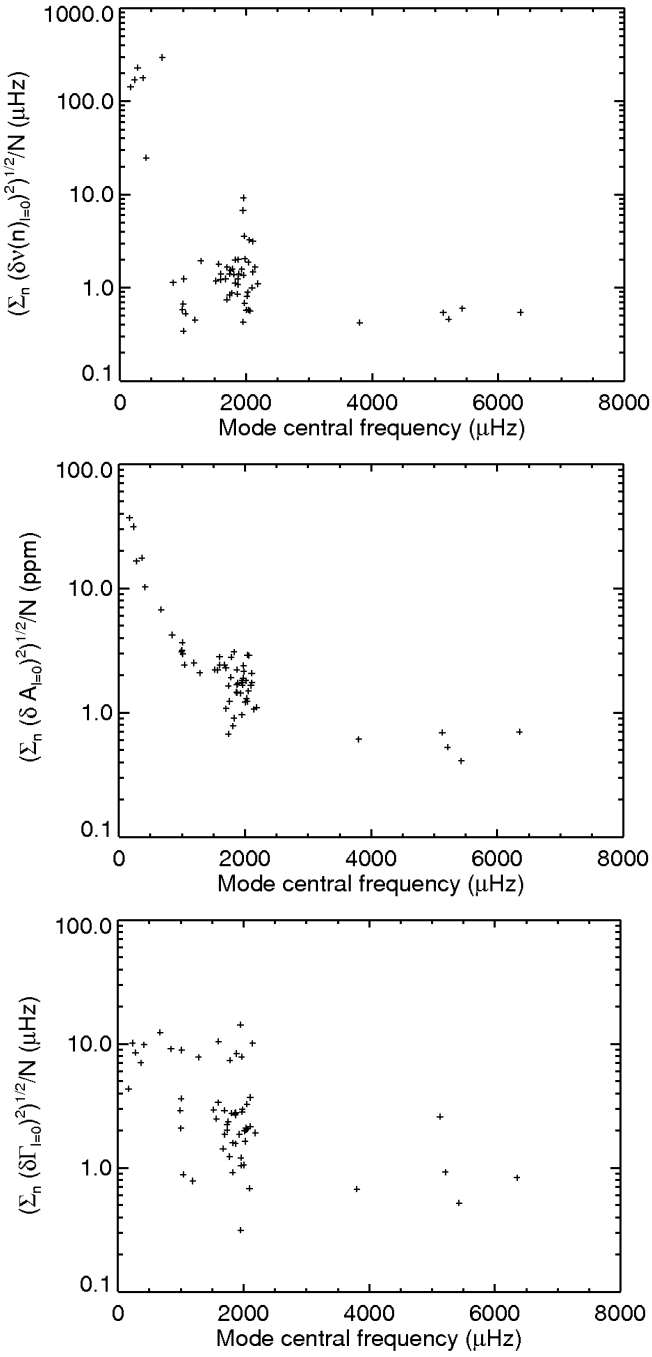


FIGURE 5.13 – Ecart quadratique moyen des erreurs pour la fréquence, l'amplitude et la largeur des modes pour chacune des étoiles simulées. En dessous de $\nu_{modes} \approx 900$ μHz, l'erreur commence à être importante : l'hypothèse de continuité des fréquences n'est plus valide. L'erreur diminue en fonction de la position de la fréquence centrale des modes ν_{modes}.

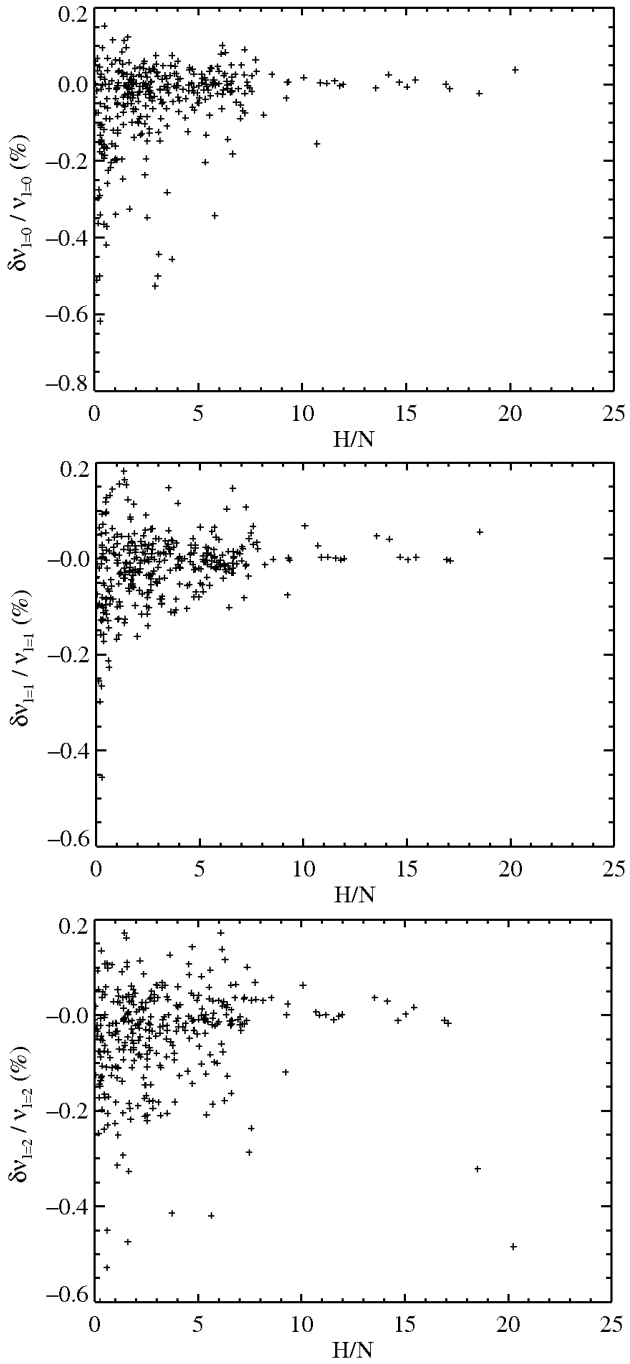


FIGURE 5.14 – Différence relative $(\nu_{n,l}^{out} - \nu_{n,l}^{in})/\nu_{n,l}^{in}$ pour les fréquences individuelles des modes après réjection des étoiles ayant une fréquence centrale $\nu_{\text{modes}} < 900$ μHz. De haut en bas, les différences relatives pour le $l = 0$, $l = 1$ et le $l = 2$. L'erreur décroit tandis que le rapport hauteur sur bruit des modes croit. Pour des rapports de hauteur sur bruit supérieurs à 10, l'erreur relative est inférieure à 0.05%, *i.e.*, inférieure à ≈ 1 μHz pour une fréquence centrale d'approximativement 2000 μHz. Cette valeur est donc très proche de la résolution spectrale.

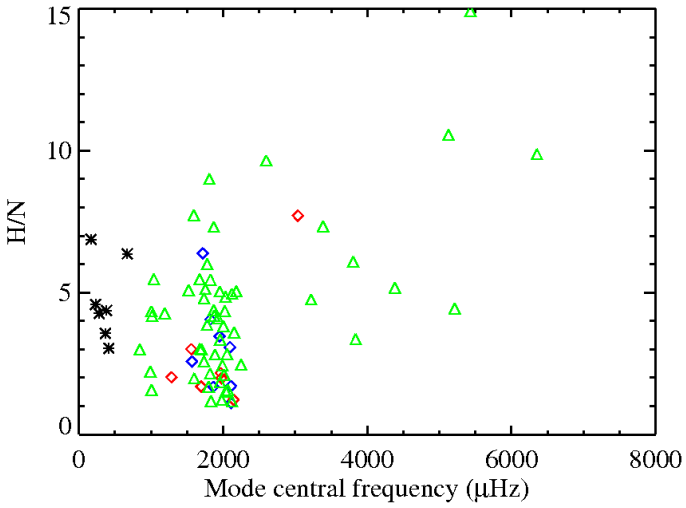


FIGURE 5.15 – Efficacité de la procédure d'identification des modes en fonction du rapport hauteur sur bruit maximum du $l = 0$ et de la fréquence centrale des modes ν_{modes}. Les étoiles noires (catégorie 4) représentent les étoiles simulées pour lesquelles l'ajustement ne fonctionne pas. Les losanges rouges (catégorie 3) regroupent les étoiles simulées pour lesquelles l'identification trouvée est incorrecte. Les losanges bleus (catégorie 2) correspondent aux étoiles simulées pour lesquelles l'identification n'est pas avérée de manière certaine. Enfin, les triangles verts (catégorie 1) représentent les étoiles pour lesquelles l'identification trouvée est la bonne. La plupart des étoiles incorrectement identifiées ont un faible rapport hauteur sur bruit.

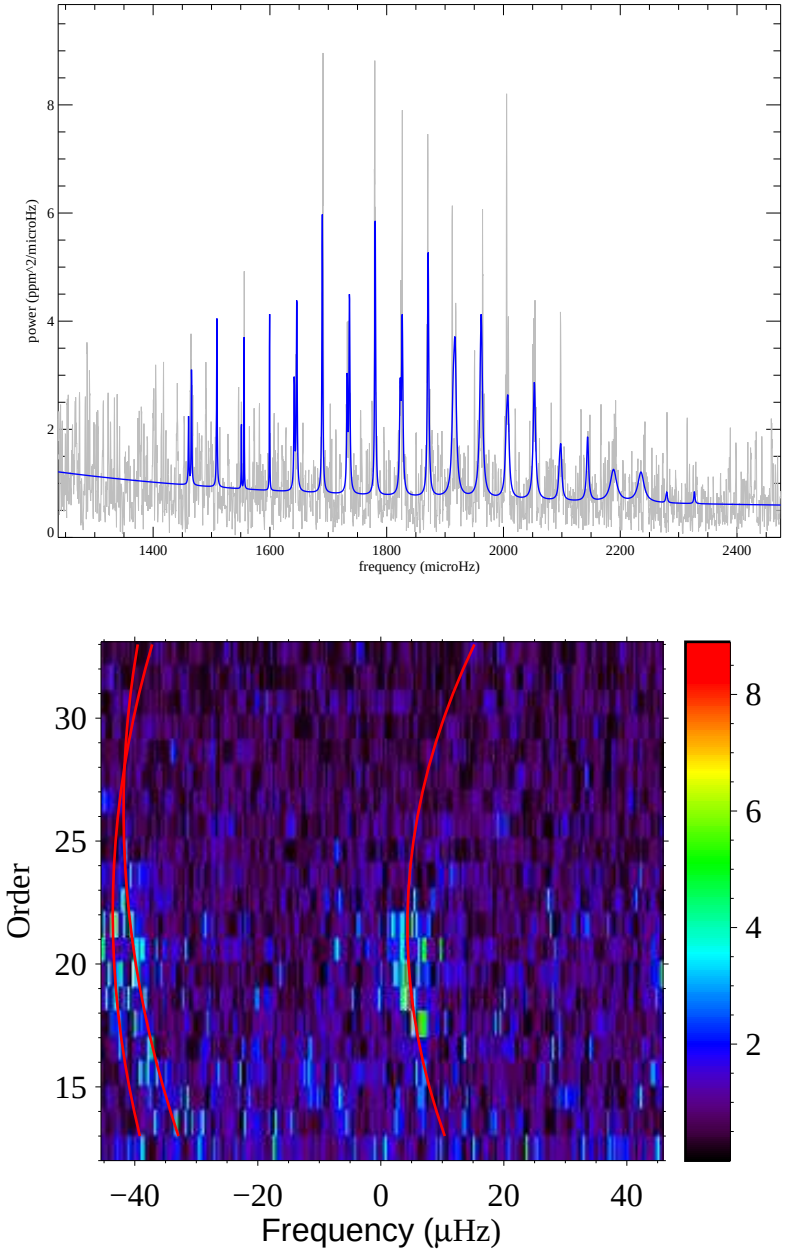


FIGURE 5.16 – Spectre de puissance et diagramme échelle obtenue pour l'une des étoiles simulées du catalogue AsteroFLAG. Sur le spectre de puissance on a représenté en ligne pleine bleue le modèle optimal trouvé. Sur le diagramme-échelle, on superpose sur le spectre le profil polynomial de distribution des fréquences. Avec une valeur maximale d'environ 7, le rapport hauteur sur bruit des modes est plutôt bon. L'identification est relativement évidente sur le diagramme échelle et est confirmée par la procédure de comparaison de modèles. L'ajustement reproduit très bien la distribution moyenne en fréquence.

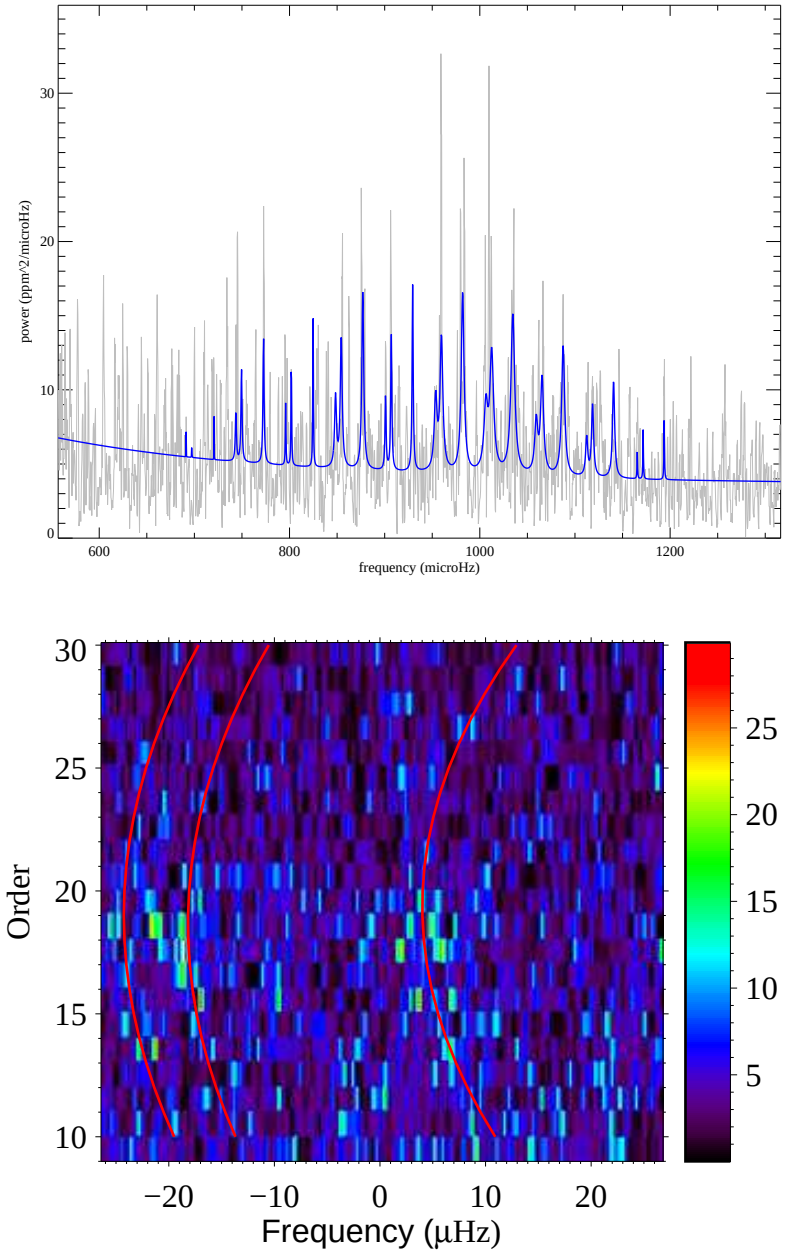


FIGURE 5.17 – Un second exemple de spectre de puissance et de diagramme échelle obtenues pour une des étoiles simulées du catalogue AsteroFLAG. Avec une valeur maximale d'environ 3, le rapport hauteur sur bruit des modes est bien moins bon que dans l'exemple de la Fig.5.16, bien qu'il y ait plus de modes ajustés. L'identification est difficile à mettre en évidence avec le diagramme échelle mais la procédure de comparaison de modèle identifie correctement les modes, en affectant une probabilité de plus de 90%. L'ajustement reproduit assez bien la distribution moyenne en fréquence, bien que l'on approche du seuil fatidique des $\nu_{modes} \approx 900$ μHz, définissant la limite à partir de laquelle l'hypothèse de continuité des fréquences commence à ne plus être valide.

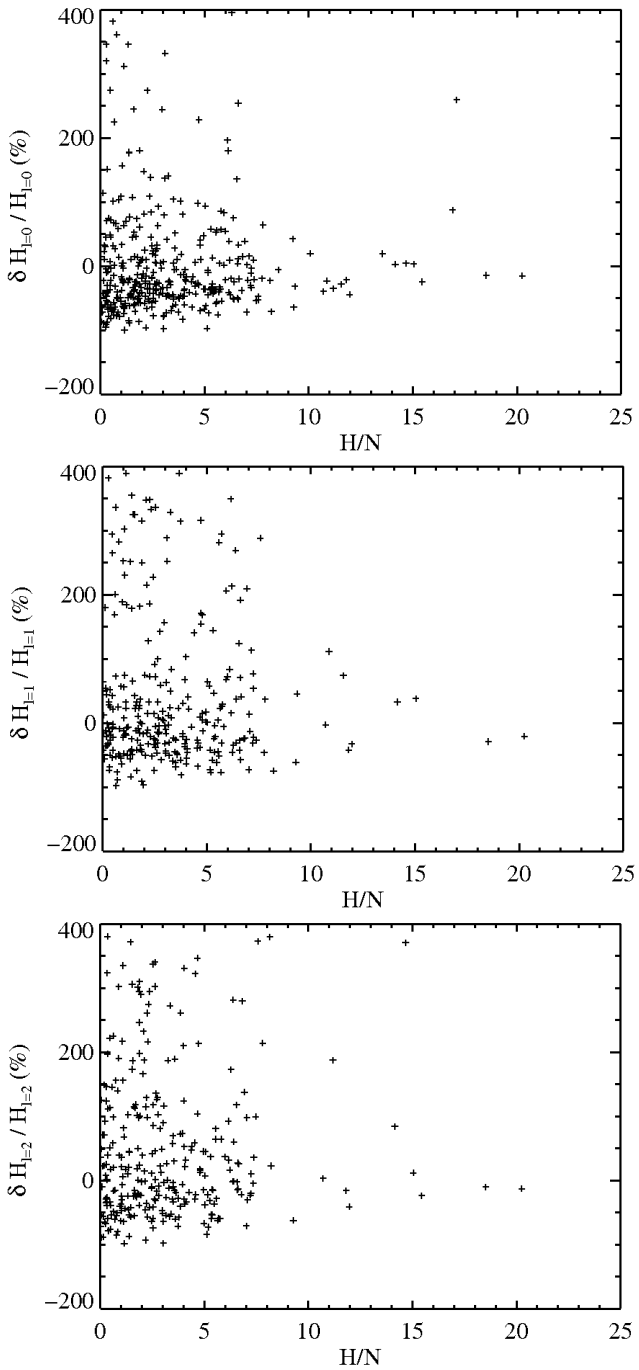


FIGURE 5.18 – Différence relative $(H^{out}_{n,l} - H^{in}_{n,l})/H^{in}_{n,l}$ pour les hauteurs individuelles pour les étoiles dont la fréquence centrale des modes ν_{modes} est supérieure à 900 μHz. De haut en bas l'erreur pour le $l = 0$, $l = 1$ et le $l = 2$. L'erreur décroit tandis que le rapport hauteur sur bruit croit. La dispersion est très prononcée pour $l > 0$ car nous négligeons les composantes rotationnelles.

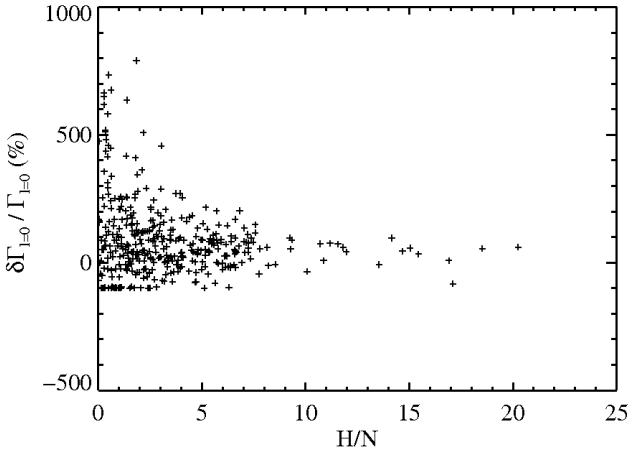


FIGURE 5.19 – Différence relative $(\Gamma^{out}_{n,l=0} - \Gamma^{in}_{n,l=0})/\Gamma^{in}_{n,l=0}$ pour la largeur des modes pour les étoiles dont la fréquence centrale des modes ν_{modes} et supérieure à 900 μHz. L'erreur décroit tandis que le rapport hauteur sur bruit croit. On rappellera que nous avons choisi d'imposer une largeur unique par ordre radial.

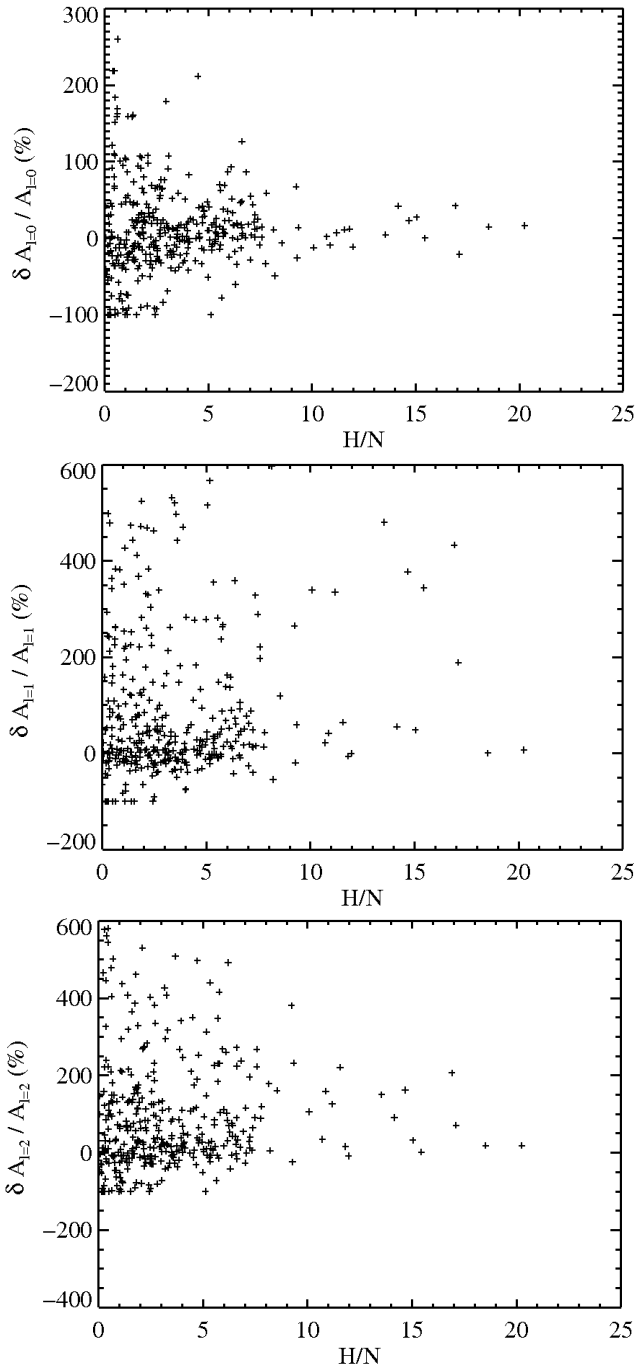


FIGURE 5.20 – Différence relative $(A^{out}_{n,l} - A^{in}_{n,l})/A^{in}_{n,l}$ pour les amplitudes des modes pour les étoiles de fréquence centrale des modes $\nu_{\max}$ supérieure à 900 μHz. De haut en bas, les erreurs pour le $l = 0$, $l = 1$ et le $l = 2$. L'erreur décroit tandis que le rapport hauteur sur bruit croit. Cependant on note une très forte dispersion pour $l > 0$ car nous négligeons les composantes rotationnelles lors de l'ajustement.

Chapitre 6

De l'analyse du signal à son interpretation physique

Sommaire

6.1 CESAM, LOSC et minimisation 144
6.2 HD49933 146
6.2.1 Physique 147
6.2.2 Méthode 147
6.2.3 Discussion des résultats 148
6.3 HD181420 153

Bien que l'extraction des paramètres individuels des modes soit une étape cruciale dans le processus d'analyse du signal stellaire, elle ne constitue pas la dernière étape : nous devons maintenant rechercher quel(s) modèle(s) stellaire(s) peut (peuvent) corroborer au mieux les observables et interpréter le(s) dit(s) modèle(s) en terme de structure interne et de propriétés physiques de l'étoile.

Cette partie de l'analyse utilisera à la fois des contraintes sismiques (fréquences) et non sismiques (luminosité, métallicité, ...). L'analyse peut se faire de bien des manières. Par exemple, il est intéressant de se baser sur des grandeurs telles que les grandes séparations et petites séparations (cf. Chap.1.2.5 pour les définitions) et leurs variations en fonction de l'ordre radial. En effet, ces grandeurs ont l'avantage d'être sensibles à des propriétés/structures différentes au sein de l'étoile.

Par exemple, les couches les plus profondes de l'étoile affectent essentiellement le $l = 1$ et le $l = 0$, car passant près du coeur. C'est ce qui explique pourquoi la petite séparation entre ces deux modes (appelée δ_{01} dans le chapitre introductif) est sensible aux propriétés du coeur de l'étoile. Les variations de la grande séparation peuvent nous informer sur les positions des discontinuités ou plus généralement, des régions de transition rapide dans l'étoile, reflétées par des changements de vitesse du son. En effet des changements de milieu se manifestent par une oscillation dans la grande séparation dont la période est directement associée à la profondeur acoustique de la discontinuité. Ceci a été montré par (Vorontsov 1988) ou encore (Monteiro et al. 1994).

Comparer les grandeurs précitées permet donc d'établir des modèles préliminaires. Une fois ces analyses préliminaires effectuées, il pourrait alors être possible de comparer directement les fréquences individuelles des modes et d'affiner la description. Cependant, il faut être prudent à cette étape car si la grande séparation n'est que peu sensible aux propriétés de surface, ce

n'est plus le cas lorsque l'on considère directement les fréquences individuelles. Or, les modèles de structure interne tels que ceux calculés CESAM2k ne décrivent que très mal la physique de surface de l'étoile. Par exemple, les écarts entre les fréquences fournies par les modèles et les données sismiques peuvent être de l'ordre de la dizaine de μHz dans le cas solaire. Il est toutefois possible d'en tenir compte en utilisant des corrections empiriques Kjeldsen et al. (2008). Parce que ma thèse est orientée préférentiellement vers l'extraction des paramètres individuels des modes, nous ne sommes pas allés jusqu'à la prise en compte des fréquences individuelles.

Pour commencer ce chapitre, nous allons présenter les caractéristiques et capacités des codes utilisés d'une part pour la modélisation de l'étoile (CESAM2k) et d'autre part pour le calcul de fréquences issues de ce modèle (LOSC). Puis nous présenterons une modélisation préliminaire de HD49933 et de HD181420.

6.1 CESAM, LOSC et minimisation

Contraindre un modèle de structure interne d'une étoile par des données sismiques nécessite de pouvoir,

1. Construire un modèle à physique donnée. C'est le rôle du code d'évolution. Dans le cadre de ce travail, nous avons utilisé le *Code d'Evolution Stellaire Adaptatif et Modulaire version 2000* (CESAM2k) (Morel 1997) et (Morel & Lebreton 2008).
2. Calculer les fréquences des modes issues de ce modèle. Ce travail a été effectué en utilisant le *Liège Oscillation Code* (LOSC) (Scuflaire et al. 2008).
3. Comparer les différentes observables aux données théoriques et établir sur la base d'un critère statistique, quel est le modèle, parmi la famille de modèles calculée, qui correspond le mieux aux observables. La statistique des erreurs systématiques étant supposées gaussienne pour les modes ajustés, nous avons utilisé un critère statistique du χ^2. Plus précisément nous avons utilisé une approche de minimisation de type Levenberg-Marquardt (LM), (Bevington & Robinson 2003).

Décrivons un peu plus chacun de ces éléments,

CESAM2k. CESAM2k est un code unidimensionnel (l'étoile est supposé à symétrie sphérique) permettant de résoudre les équations de la structure interne décrites Chap.1.2.2. Il se différencie de la plupart des autres codes numériques par sa flexibilité tant au niveau de la précision numérique qu'au niveau de la physique qu'il est capable de prendre en compte.

Par exemple, le découpage en modules autonomes rend possible de choisir la physique à prendre en compte très facilement : diffusion microscopique, pénétration de la zone convective dans les zones radiatives (*overshoot*), théorie de la convection, table d'opacités,...

Lors du calcul du modèle de structure interne, l'étoile évolue d'un modèle initial sur la ZAMS jusqu'à ce que l'âge final demandé soit atteint [1]. La précision du calcul peut être modifiée en jouant sur une vaste gamme de paramètres (on en compte plus d'une vingtaine). Par exemple, il est possible de modifier les valeurs du pas de temps minimal et maximal à utiliser lors du calcul. Ce paramètre peut aider lorsque l'on constate, sur les modèles par exemple, un effet mémoire durant la rétractation du coeur convectif. Ou encore, on peut modifier le nombre de couches maximum constituant l'étoile. Ceci peut permettre de résoudre certains problèmes numériques dans les régions de variation rapide de la structure interne.

De par l'étendue des capacités de CESAM, il serait inutile de le décrire dans son intégralité. Au lecteur intéressé par les méthodes numériques utilisées et par une liste des phénomènes

1. D'autres conditions peuvent aussi être utilisés pour arrêter le calcul, tel que la température effective.

physiques traités à ce jour par le code, nous recommandons de se référer à la documentation (Morel 2009). Dans la suite, nous ne citerons que quelques particularités intéressantes.

Régulièrement mis à jour, CESAM2k prend en compte la physique la plus récente et peut être appliqué à n'importe quelle phase évolutive (PMS, séquence principale -MS-, phase des sous-géantes et des géantes rouges). Au niveau des abondances chimiques, le nombre de cycles de réaction nucléaire pris en compte peut être modifié. En ce qui nous concerne, nous n'avons tenu compte que des chaînes PP et du cycle CNO. Ce choix est amplement suffisant car les objets analysés jusque là on une masse relativement faible (inférieure à $1.5M_{\odot}$) et sont très probablement dans la MS. Les abondances chimiques peuvent être basées sur les abondances solaires fournies par Grevesse & Noels (1993) aussi noté GN93, ou Asplund et al. (2005).

Parmi les processus physiques pris en compte par CESAM2k, il y en a deux qu'il est particulièrement utile de décrire.

Tout d'abord l'overshooting instantané. En étendant le mélange au-delà de la frontière de Schwarzschild, il a pour effet d'étendre la zone de mélange. Il permet aussi de lisser la discontinuité entre les deux milieux en fixant dans la zone radiative, le gradient de température égal au gradient adiabatique[2]. L'étendue de la pénétration convective s'exprime en hauteur de pression : $d_{ov} = \alpha_{ov} H_p$ (dans le coeur : $d_{ov} = \alpha_{ov}\, min(r_c, H_p)$ où r_c représente le rayon du coeur), avec α_{ov} le coefficient d'overshoot. Notons qu'il est aussi possible d'utiliser d'autres définitions, tel que celle de Zahn (1991). Notons aussi que l'overshoot est ici un paramètre *ad hoc* pouvant être mimé par des processus physiques telle que le mélange rotationnel (en cours d'implémentation dans CESAM2k) : l'overshoot n'est que le reflet de notre incapacité à correctement décrire les phénomènes convectifs et principalement l'inexactitude du critère de Schwarzschild (définissant la taille de la zone convective).

CESAM2k est capable de gérer la diffusion des éléments chimiques. La diffusion microscopique a un rôle fondamental pour expliquer l'abondance d'hélium en surface, car la diffusion des éléments lourds et de l'hélium vers le coeur engendre un abaissement notable de la métallicité de surface. Pour le Soleil par exemple, la fraction de ^4He de surface baisse d'environ 10% entre la ZAMS et l'âge actuel. Dans les zones convectives ce processus de diffusion rentre donc en competition avec la convection. La diffusion à l'echelle microscopique peut être en fait séparée en trois composantes compétitives,

La diffusion thermique. En présence d'inhomogénéités de matière, de température ou de pression, elle tend à réduire les gradients de concentrations (homogénéisation).

La diffusion gravitationnelle. Elle est aussi appelée sédimentation gravitationnelle (*gravitational settling*). La gravité provoque une migration des éléments les plus lourds vers le coeur de l'étoile, ce qui engendre une stratification des espèces chimiques.

L'accélération radiative. Elle contrecarre les effets de la diffusion gravitationnelle car la pression radiative pousse les éléments à remonter en surface. Sa magnitude est essentiellement liée à l'opacité et à la température du milieu. Dans le cas solaire, cet effet est négligeable, cependant il peut devenir significatif à partir d'environ $1.5M_{\odot}$. Nous n'en avons pas tenu compte dans nos modélisations.

D'une manière générale, la diffusion microscopique d'une particule s'exprime dans CESAM2k selon Iben & MacDonald (1985) et Burgers (1969).

LOSC. Le code d'oscillation a pour rôle important de déterminer les fréquences d'oscillation de l'étoile. Pour cela, LOSC (Scuflaire et al. 2008) suppose les oscillations adiabatiques et nécessite

2. Certains codes ne prennent pas le gradient de température égal au gradient adiabatique mais utilisent le gradient radiatif et étendent simplement le mélange.

en entrée les variables de rayon, masse, de pression, densité et l'exposant adiabatique. Bien entendu, supposer les oscillations adiabatiques affectera la détermination des fréquences les plus sensibles aux effets de surface de l'étoile. En surface par exemple, les oscillations ne sont justement pas adiabatiques. Pour le Soleil, ces erreurs peuvent aller jusqu'à une dizaine de μHz à haute fréquence.

Approche de Levenberg-Marquardt. L'algorithme de Levenberg-Marquardt est spécifiquement conçu pour fonctionner avec un critère des moindres carrés (χ^2) dont la fonction de vraisemblance présuppose les erreurs gaussiennes. La fonction coût associée s'écrit,

$$\chi^2(\boldsymbol{x}) = \frac{1}{2}\sum_i \frac{(\boldsymbol{x_i^{(th)}} - \boldsymbol{x_i^{(obs)}})^2}{\sigma_i^2}, \tag{6.1}$$

où $\boldsymbol{x^{theo}}$ et $\boldsymbol{x^{obs}}$ correspondent respectivement aux paramètres théoriques décrivant les modèles et à leur équivalent observationnel. σ correspond à l'erreur sur la donnée observationnelle.

C'est une combinaison entre l'algorithme de recherche de la plus grande pente (*gradient descent*) et de Gauss-Newton (basé sur les dérivées secondes). Comme tout algorithme de recherche de minimum local, il est sensible à la position initiale dans l'espace des paramètres. En supposant qu'il existe un unique maximum dans la fonction erreur définie par le χ^2, un algorithme de Levenberg-Marquardt est plus rapide et précis que les alternatives existantes. Il se formule,

$$\boldsymbol{x_i} = \boldsymbol{x_{i-1}} - \boldsymbol{H} + \lambda diag(\boldsymbol{H})^{-1} d, \tag{6.2}$$

tel que d et $\boldsymbol{H}$ sont respectivement une estimation du gradient et de la hessienne. λ est un coefficient de couplage et $\boldsymbol{x}$ correspond aux vecteurs de paramètres. On peut montrer que lorsque λ est grand, l'algorithme se comporte comme une descente de gradient tandis que lorsque λ est petit, il suit plutôt un schéma de Gauss-Newton. λ est modifié à chaque itération de telle sorte à privilégier la descente de gradient lorsque celui ci est fort (domaine non linéaire) et plutôt la méthode Gauss-Newton si le gradient est faible (région en plateau).

Le couplage de ces 3 éléments est effectué par un programme FORTRAN, initialement écrit par Marc-Antoine Dupret. Nous l'avons quelque peu adapté à la méthodologie que nous avons choisie.

Notons enfin qu'il serait préférable d'utiliser un code MCMC pour effectuer la minimisation car les corrélations entre paramètres stellaires (masse, âge, abondances, overshoot, diffusion, ...) sont très forts et les contraintes à notre disposition plutôt réduite. Dans de telles conditions on s'attend à de multiples maxima de probabilité (respectivement de multiples minima du χ^2) et l'algorithme LM peut rencontrer des difficultés. Cependant, en tant que travail préliminaire nous n'avons pas jugé nécessaire d'établir un échantillonnage car la barrière temporelle est le principal obstacle à l'utilisation des MCMC dans les codes stellaires. Ce point sera discuté ultérieurement, dans le chaptire de conclusion.

6.2 HD49933

L'analyse du spectre de HD49933 s'est révélée riche en information : pas moins de 18 ordres radiaux ont pu être extraits. Toutefois, les barres d'erreurs associées sont variables : le $l = 1$ sont très bien contraint car non perturbé par la présence du $l = 2$ (le $l = 3$ étant négligeable) tandis que le $l = 0$ est entaché d'erreurs bien plus grandes. Grâce à l'utilisation du LRa-01 nous sommes maintenant en mesure d'utiliser les fréquences extraites pour établir une modélisation de l'étoile.

6.2.1 Physique

Le calcul a été effectué en utilisant l'équation d'état OPAL 2005 (Rogers & Nayfonov 2002). Les tables d'opacités sont celles de Rogers & Iglesias (1996) et Alexander & Ferguson (1994) pour les hautes/basses températures et on utilise la mixture GN93. L'atmosphère de l'étoile est supposée suivre l'approximation d'Eddington, c'est à dire que le mécanisme d'échange est essentiellement radiatif proche de la surface. Cette hypothèse est incorrecte car les couches externes d'une étoile de type solaire sont convectives. Toutefois un artifice numérique (cf. Morel (2009) page 111) permet d'obtenir une description équivalente à celle donnée par la MLT si la profondeur optique est suffisamment grande (*i.e.* $\tau >> 1$). Nous supposerons que nous sommes dans cette situation.

Les taux de réaction nucléaires ont été récemment revus pour certaines réactions. Joao-Pedro Marques a implémenté les taux de réaction révisés (dit LUNA, du nom de l'expérience : *Laboratory for Underground Nuclear Astrophysics*) pour l'une des réactions prépondérantes du cycle CNO : la réaction azote + proton, notée $^{14}N(p,\gamma)^{15}O$ (Formicola et al. 2004). Les nouveaux taux sont environ 60% plus faibles que les taux NACRE (pour *Nuclear Reaction Rates for Charged-particle REactions*) habituellement utilisés. La nature convective du coeur est étroitement associée à la température au coeur. Lorsque les températures sont suffisantes le cycle CNO peut contribuer significativement au terme de production d'energie. Or la réaction $^{14}N(p,\gamma)^{15}O$ est la plus lente de ce cycle. Son taux affectera donc indirectement l'âge de l'étoile et l'extension de l'eventuelle zone convective du coeur. Pour tenir compte au mieux des phénomènes convectifs, nous avons choisi d'utiliser le formalisme CGM. Comme expliqué Chap.1, ce formalisme tient compte de la cascade turbulente, contrairement au MLT : avec une masse voisine de $1.2M_\odot$ il est probable que HD49933 dispose d'un coeur convectif, mais aussi d'une envelope convective plus fine que le Soleil.

6.2.2 Méthode

Nous avons essayé de rechercher un ajustement idéal des observables à notre disposition en faisant varier la masse $M/M_\odot$, l'âge, la métallicité initiale de l'étoile $(Z/X)_0$, l'abondance initiale en hélium Y_0, le coefficient de convection α_{CGM} et le coefficient d'overshoot α_{ov}.

Concernant les observables sismiques, la méthode employée consiste à comparer certaines propriétés globales de la grande séparation $\Delta\nu$ et de la petite séparation δ_{01}. Par exemple, au premier ordre, nous pouvons établir une comparaison des valeurs moyennes de ces grandeurs. Cependant, comme montré sur la Fig.6.1 les variations constatées des dits paramètres sismiques ne sont pas uniquement dû aux incertitudes de mesures : la valeur moyenne est insuffisante. C'est le cas de la grande séparation par exemple, pour laquelle la somme des écarts quadratiques entre la moyenne et les données est supérieure à 9, (dû à l'oscillation de celle ci, interprétable physiquement).

Nous avons donc utilisé une représentation plus fine : $\Delta\nu$ est représentée par une fonction oscillante et une constante,

$$\Delta\nu^*(\nu) = C_0 + A_0 sin(\omega\nu + \phi). \tag{6.3}$$

La petite séparation $\delta\nu_{01}$ est approximée par,

$$\delta\nu_{01}(\nu) = a + b\nu. \tag{6.4}$$

Autrement dit, nous avons calculé un premier ajustement (et les barres d'erreurs correspondantes) de ces grandeurs sismiques sur la base des observations sismiques (fréquences) à notre disposition (étape de réduction des données). Cette étape est effectuée par une recherche de la plus grande

pente (*Gradient descent* en anglais) dans un contexte MLE. Ensuite, les paramètres C_0, A_0, a, b ont été comparés à leurs homologues théoriques, calculés après chaque itération de l'algorithme de LM (étape de minimisation du χ^2). Dans les deux étapes on utilise conjointement les définitions centrées sur $l = 0$ et sur $l = 1$ (par exemple $\delta\nu_{01}^{(0)}$ et $\delta\nu_{01}^{(1)}$ pour la petite séparation). Notons toutefois que les corrélations entre variables n'ont pas été prises en compte car le code originel à notre disposition n'est pas capable d'en tenir compte.

Cette approche a l'avantage de réduire la quantité de contraintes. En effet nous avons constaté qu'utiliser directement les valeurs individuelles des contraintes sismiques $\Delta\nu$ et $\delta\nu_{01}$ rend la convergence vers les véritables maxima de probabilité assez difficile : il semble exister un très grand nombre de minima locaux dans le χ^2 rendant la solution extrêmement sensible aux conditions initiales avec l'algorithme actuel. Le problème le plus fondamental vient en fait de la valeur des fréquences issues d'un modèle. En effet, les effets de surface pouvant être importants (et mal caractérisés lors de la modélisation) l'ordre radial n est souvent connu à ± 1 près : dans ce cas comment établir la différence $\nu^{obs}(n,l) - \nu^{theo}(n,l)$ au coeur du calcul du χ^2 sans une connaissance préalable assez fine des paramètres stellaires (âge, masse, abondances,...) ?

En plus des observables sismiques, nous avons utilisé la luminosité $log(L/L_\odot)$, la température de surface T_{eff} et la métallicité de surface $(Z/X)_S$. Cette dernière est calculée sur la base de la métallicité solaire à mixture donnée $(Z/X)_\odot$ via la relation,

$$\left(\frac{Z}{X}\right)_S = \left(\frac{Z}{X}\right)_\odot 10^{-[Fe/H]_S}. \tag{6.5}$$

Pour GN93 on a $(Z/X)_\odot = 0.0245 \pm 0.0027$.

Les analyses non sismiques de HD 49933 sont diverses et variées. Selon les travaux, la température de surface peut varier de 6750 ± 130 (Bruntt et al. 2008) à 6550 ± 100 K (Bruntt 2009). Selon Michel et al. (2008) on a $log(L/L_\odot) = 0.54 \pm 0.02$. Enfin, la métallicité est donné par Gillon & Magain (2006), $[Fe/H]_S = -0.37 \pm 0.03\, dex$.

La recherche des paramètres se fait dans un espace délimité par des *a priori* uniformes, dont les valeurs caractéristiques sont résumées Table 6.2. L'objectif de ces *a priori* est d'éviter de chercher des solutions peu crédibles, comme par exemple, des abondances en hélium initiales inférieures à l'abondance primordiale de l'univers.

6.2.3 Discussion des résultats

La recherche du minimum du χ^2 a été effectuée en prenant le meilleur χ^2 obtenu parmi un ensemble de minimisations différant par leur vecteur initial de paramètres. Dans cette section nous résumons les principaux résultats obtenus. Notamment, on retrouvera Table 6.1 les paramètres stellaires fondamentaux obtenus après minimisation. Les Tables 6.2 et Table 6.3 résument quant à elles les *a priori* et les contraintes utilisées.

Le critère retenu pour identifier les interfaces entre zone convective et radiative est le terme A fourni par CESAM2k, défini comme,

$$A = \frac{1}{\Gamma_1}\frac{\partial lnP}{\partial lnR} - \frac{\partial ln\rho}{\partial lnR}. \tag{6.6}$$

Ce critère est en fait proportionnel à la fonction de Brunt-Vaisala puisqu'en se référant à l'Eq.1.22,

$$N^2 = \frac{g}{R}A. \tag{6.7}$$

	Ni overshoot ni diffusion	Overshoot seul	Overshoot + diffusion
$Age\ (Mans)$	3.516 ± 0.920	3.257 ± 0.168	3.208 ± 0.033
$Masse\ (M_\odot)$	1.13 ± 0.09	1.18 ± 0.08	1.19 ± 0.07
$R/R_\odot$	1.401	1.425	1.406
$(Z/X)_0$	0.0081 ± 0.0011	0.0106 ± 0.0010	0.0100 ± 0.0006
Y_0	0.259 ± 0.03	0.257 ± 0.043	0.248 ± 0.010
Xc	0.03	0.40	0.43
α_{CGM}	0.60 ± 0.07	0.60 ± 0.04	0.67 ± 0.05
α_{ov}	–	0.27 ± 0.09	0.25 ± 0.05
$L/L_\odot$	0.534	0.522	0.520
T_{eff}	6640	6540	6578
Epaiss. $Z_C^{inf}(\%R_{star})$	–	8.8%	8.9%
Epaiss. $Z_C^{sup}(\%R_{star})$	10.0%	10.5%	11.2%
χ^2	10.9	5.9	33.9

TABLE 6.1 – **HD49933.** Principaux résultats après minimisation pour les 3 cas envisagés. Les deux avant dernières lignes indiquent l'extension des zones convectives en fraction de rayon stellaire.

Les premiers modèles calculés n'ont pas fait intervenir l'overshoot. Avec la physique décrite précédemment, nous avons constaté alors qu'il n'est en aucun cas possible de décrire correctement à la fois la petite et la grande séparation. La structure interne de l'étoile obtenue pour le meilleur ajustement est présentée Fig.6.1 (ligne bleu pleine). Comme on peut le voir, la structure se caractérise essentiellement par une zone convective externe relativement fine et l'absence de zone convective au coeur. L'étoile est en fin de séquence principale avec une abondance en hydrogène au coeur quasi-nulle.

Pour obtenir des modèles satisfaisant les deux critères sismiques utilisés il nous faut faire intervenir l'overshoot. Ainsi, Fig.6.1 on remarquera que le meilleur modèle présente cette fois un coeur convectif relativement petit sans que, globalement la structure des couches supérieures soit changées. Ceci confirme que la contrainte $\delta\nu_{01}$ en moyenne, caractérise bien les phénomènes au coeur mais peu les couches supérieures (cf. Chap.1.2.5). L'abondance en hydrogène au coeur est significativement plus forte comparativement au modèle sans overshoot : l'étoile est approximativement à mi-vie cette fois ci. Le χ^2 est aussi meilleur. Comparativement, les modifications de χ^2 sont majoritairement dues à $\delta\nu_{01}$ car même si T_{eff} est mieux reproduite avec overshooting, la luminosité s'éloigne plus de la contrainte imposée (effets compensatoires).

Finalement, nous pouvons conclure qu'il est crucial d'incorporer l'overshoot à partir du coeur pour modéliser relativement bien l'étoile : celui-ci a une importance cruciale sur la position de l'étoile dans son cycle de vie car il amène plus facilement de l'hydrogène des couches supérieures vers le coeur, prolongeant la vie de l'étoile.

Après observation de l'ajustement de $\delta\nu_{01}$, Fig.6.1 on constate toutefois que même si l'ajustement est bon, la pente est assez mal reproduite. Nous pensons qu'un éventuel mécanisme physique en jeu n'est pas pris en compte ici. Dans le soleil, la diffusion microscopique améliore la modélisation, en conséquence on a cherché si cela pouvait être le cas dans HD49933.

Les résultats de l'ajustement des contraintes sismiques sont résumés Fig.6.2 (trait pointillé épais) et les Fig.6.5 représentent quelques paramètres de la structure interne pour ce dernier cas. L'incorporation de la diffusion microscopique a rendu les temps de calculs particulièrement longs et la dépendance envers les paramètres initiaux n'a pas été testée de manière aussi intensive

	$Age\ (Mans)$	$Masse\ (M_\odot)$	$(Z/X)_0$
a priori	$U[1000, 6000]$	$U[0.9, 1.5]$	$U[0.006 - 0.015]$
	Y_0	α_{CGM}	α_{ov}
a priori	$U[0.245, 0.300]$	$U[0.4, 1.0]$	$U[0.0, 0.6]$

TABLE 6.2 – **HD49933.** Plages tolérées lors de la recherche du meilleur modèle stellaire.

$T_{eff}\ (K)$	$log(L/L_\odot)$	$(Z/X)_S$	$C_0\ (\mu Hz)$	$A_0\ (\mu Hz)$
$6550 \pm 100.$	0.54 ± 0.02	0.0105 ± 0.003	85.5 ± 0.5	1.39 ± 0.1
$\omega\ (rad.ks)$	ϕ	$a\ (\mu Hz)$	b	
13 ± 2	1.7 ± 1.4	4 ± 1	-1.3 ± 0.5	

TABLE 6.3 – **HD49933.** Résumé des contraintes observationnelles appliquées lors de la recherche du meilleur modèle stellaire. Toutes les contraintes sont gaussiennes car nous effectuons une minimisation du χ^2.

que les deux cas précédents. De plus, certains modèles semblent converger assez mal et subissent quelques aléas numériques malgré nos tentatives d'affinement de la structure interne (nombre de couches maximum, pas temporel minimum/maximum, constante de répartition "psi0" de CESAM2k). Ces résultats préliminaires sont donc à prendre avec précaution. Globalement la structure interne de l'étoile est inchangée : le coeur convectif et la zone convective ont approximativement la même extension qu'avec seulement l'overshooting. La différence notable vient de l'abondance en hélium de surface : la diffusion contrecarre la convection en provoquant une migration de l'hélium de la zone convective externe vers les couches plus profondes. En conséquence, l'abondance de l'hélium en surface est inférieure à l'abondance primordiale. En soi cela reste cohérent : cet effet est exactement celui attendu et constaté dans le cas solaire. Ici, comme l'abondance d'hélium initiale obtenue pour HD 49933 est déjà particulièrement faible (autour de 0.25), la quantité de surface est bien plus faible que pour le Soleil : 0.15 au lieu de 0.245.

Enfin, l'absence de prise en compte de la force radiative et de la diffusion turbulente peut aussi jouer sur cette valeur de surface car ces deux phénomène homogénéisent les abondances.

Le χ^2 obtenu est plutôt mauvais. A l'issue d'une comparaison avec le modèle prenant en compte l'overshooting, on constate que celui ci est essentiellement pénalisé par C_0 (à environ 2σ de la valeur moyenne : 86.6 au lieu de 85.5 ± 0.5 μHz) et surtout par l'amplitude de l'oscillation (à plus de 4σ de la valeur imposée). La petite séparation n'a pas été modifiée par l'introduction de la diffusion microscopique : les conséquences se font essentiellement sentir en surface. Enfin, on remarquera que les barres d'erreurs sur des paramètres comme l'âge ou la métallicité initiale sont anormalement petites.

A première vue, on pourrait conclure que la diffusion n'est pas un facteur à prendre en compte si la physique est correct. Toutefois, à cause des difficultés explicitées dans les paragraphes précédents (difficulté de convergence, présence très probable de minima locaux et prise en compte de phénomènes physiques alternatifs tels que la diffusion turbulente) nous devons rester prudent : de notre point de vue, il n'est ni exclu ni avéré que la diffusion microscopique améliore la modélisation.

Pour conclure nous pouvons affirmer avec une très grande confiance que HD49933 est une étoile à mi-vie dont la structure interne est constituée à la fois d'un coeur convectif et d'une zone convective de surface, relativement fine, s'étendant au total sur environ 20% du rayon de l'étoile. La masse obtenue ($1.18 M_\odot$) semble compatible avec les précédentes estimations ($1.20 M_\odot$ selon

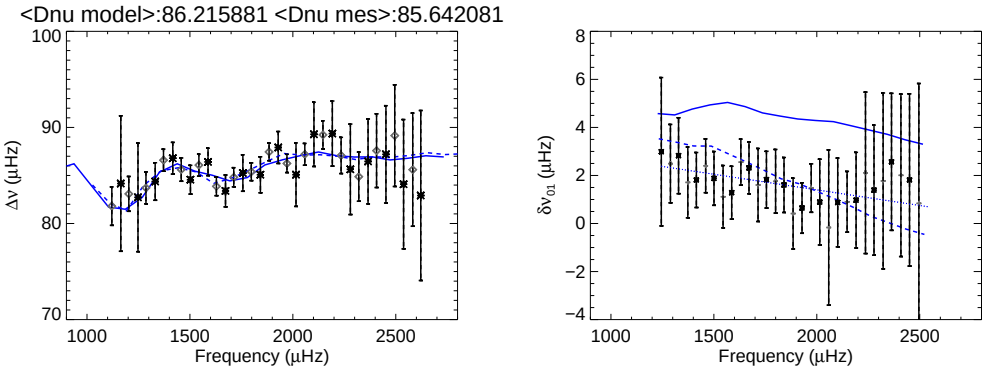


FIGURE 6.1 – **HD49933.** Comparaison entre les modèles obtenus sans diffusion microscopique : avec overshooting (ligne en pointillé épais) ou sans overshooting à partir du coeur (ligne pleine). A gauche, pour la grande séparation $\Delta\nu$, à droite pour $\delta\nu_{01}$. Pour ce dernier on a rajouté en pointillés fins le résultat de l'ajustement préliminaire des données par un polynôme d'ordre 1. Bien que le résultat sur les paramètres stellaires ajustés dépende du vecteur de paramètre initial, il n'existe aucun modèle sans overshooting capable de bien reproduire la petite séparation.

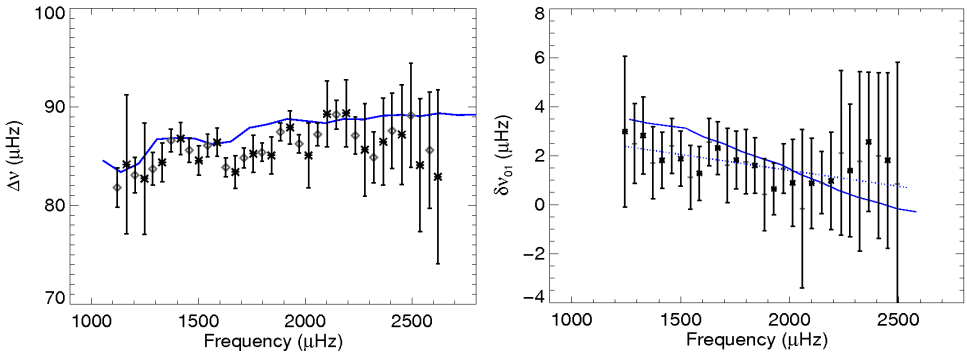


FIGURE 6.2 – Même chose que les figures précédentes, avec diffusion microscopique. Il n'existe pas d'effet significatifs de la diffusion microscopique sur les modèles. On ne peut exclure un éventuel impact de celle-ci. On soulignera encore une fois que la présence de minima locaux peut modifier l'ajustement.

Mosser et al. (2005)) pour un rayon d'environ $1.42R_\odot$.

On fera remarquer que tout comme la grande séparation, la petite séparation subit des fluctuations qui semblent périodiques. Comme signalé par Roxburgh (2009), même si la tendance moyenne de la petite séparation reflète essentiellement les propriétés au coeur de l'étoile, elle est aussi affectée partiellement par les effets des couches les plus externes. Ainsi sa période d'oscillation est reliée au rayon acoustique de l'enveloppe convective externe. Ces oscillations sont donc probablement en partie dues à des phénomènes sans rapport avec l'intérieur stellaire profond. Toutefois, il est étonnant d'avoir d'aussi grandes amplitudes. En comparaison, la petite séparation obtenue théoriquement pour les modèles ne subit pas de variations si importantes. Aucun modèle d'intérieur stellaire que nous avons calculé n'arrive à reproduire ce genre d'oscillation. Dans le Soleil, les oscillations sont de faibles amplitude. En résumé, selon nous, il est possible que (classé du moins probable au plus probable),

1. la corrélation mutuelle entre valeurs voisines de la grande et des petites séparations pro-

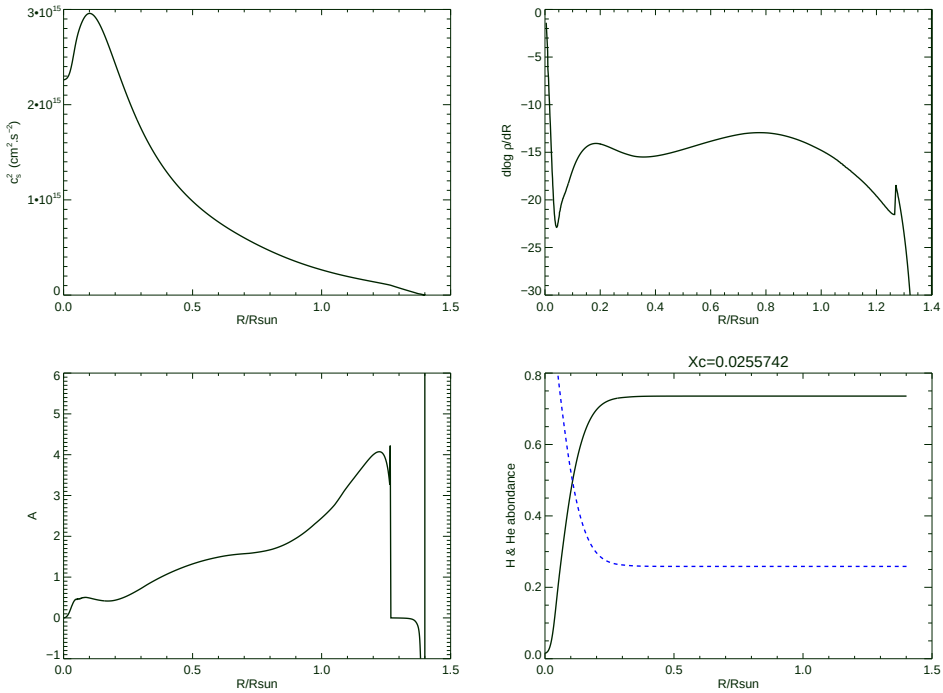


FIGURE 6.3 – **HD49933.** Vitesse du son, dérivée première de la densité, terme A et abondances en hydrogène et hélium en fonction du rayon de l'étoile (exprimé en rayon solaire) dans le scénario sans overshoot ni diffusion. On remarque qu'il n'existe peu ou pas du tout de coeur convectif dans ce modèle. L'étoile quitte la séquence principale puisque l'abondance d'hydrogène au coeur est nulle. Ce modèle est très peu probant d'une part à cause du χ^2 élevé et d'autre part à cause de son incapacité à reproduire correctement la contrainte $\delta\nu_{01}$.

voque artificiellement cette oscillation. Ce sont alors des fluctuations dues au bruit que nous observons.

2. le phénomène d'oscillation est dû à un comportement physiquement explicable, mais la physique implémenté dans CESAM2k n'est pas capable de reproduire.
3. du fait de la procédure de minimisation qui ne recherche qu'un minimum local et qui utilise une modélisation sommaire des fréquences (polynôme d'ordre 1), nous avons raté le modèle adéquat capable de reproduire ces oscillations.

Enfin soulignons que cette approche nécessite d'être complétée par une analyse des dépendances à la composition chimique utilisée. Par exemple pour le Soleil, nous savons que les tables d'opacités de AS05 modifie significativement la structure interne de notre étoile au regard de la qualité des données à disposition. Il serait aussi intéressant de voir à quel point le mélange turbulent peut améliorer/dégrader la qualité de l'ajustement de l'ensemble des contraintes observationnelles. D'autant plus qu'avec une période de rotation de l'ordre 3 à 4 jours, celle ci pourrait avoir un effet majeur sur le transport de chaleur et le mélange d'espèces chimiques. En effet, la rotation induit une circulation méridienne (advection) et, très probablement un mélange convectif turbulent des espèces chimiques.

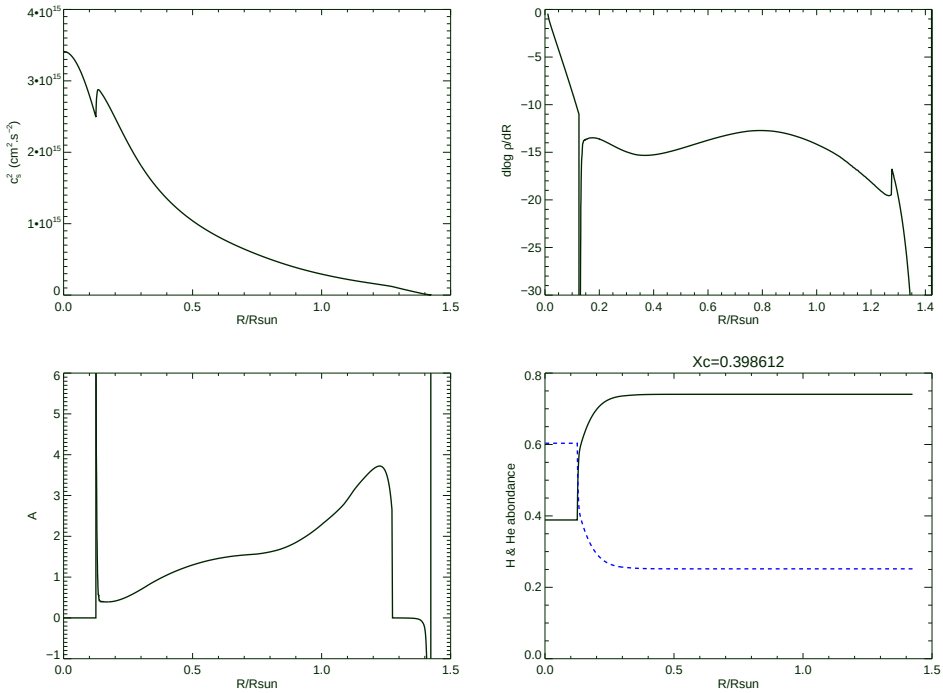


FIGURE 6.4 – **HD49933.** Vitesse du son, dérivée première de la densité, terme A et abondances en hydrogène et hélium en fonction du rayon de l'étoile dans le scénario avec overshoot et sans diffusion. Un coeur convectif représentant est apparu et la position de la zone convective externe n'est pas radicalement modifiée comparativement à la Fig.6.3 (sans overshoot). On fera remarquer que l'étoile est cette fois à mi-vie environ.

6.3 HD181420

La procédure suivie pour ajuster les données de HD181420 est très similaire à celle utilisée pour HD49933 : on recherche à ajuster la grande séparation, modélisée par une fonction sinusoïdale et une constante. La petite séparation est modélisée par un fonction linéaire. Sur la Fig.6.6 on a représenté une comparaison du résultat de cette étape de réduction de données avec les observables sismiques précitées. On remarquera que cette modélisation ne reproduit que très mal le profil des fréquences autour de 1500 μHz. Dans le Chap.4.3.2, nous avions signalé qu'un mode $l = 1$ ne semblait pas suivre l'oscillation périodique de ses voisins sur le diagramme échelle sur la plage $[1000, 2000]$ μHz. La grande séparation, basée sur des différences de fréquences est naturellement impactée : on voit une cassure dans le profil de son oscillation. Ce problème n'affecte tout au plus que 2 modes mitoyens de même degré à cause des corrélations introduites lors du calcul de la grande séparation. Le "défaut" d'oscillation de la grande séparation semble aussi affecter un ou quelque modes de degré $l = 0$ (voir par exemple Fig.6.6), ce qui discrédite en partie l'hypothèse formulée dans le Chap.4.3.2, incriminant la nature mixte du mode $l = 1$ à 1385.18 μHz. Nous devons garder à l'esprit que les pics orbitaux ont éventuellement introduit des biais [3]. Quoiqu'il en soit, la modélisation que nous proposons permet tout de même de tenir

3. Mais comme signalé dans le chapitre traitant de l'analyse de HD181420, il est peu probable que ce biais soit important (si tant est qu'il y ait biais) car les pics orbitaux n'affectent en principe que quelques points du spectre ; points qu'on a pris la précaution d'enlever avant l'ajustement des modes.

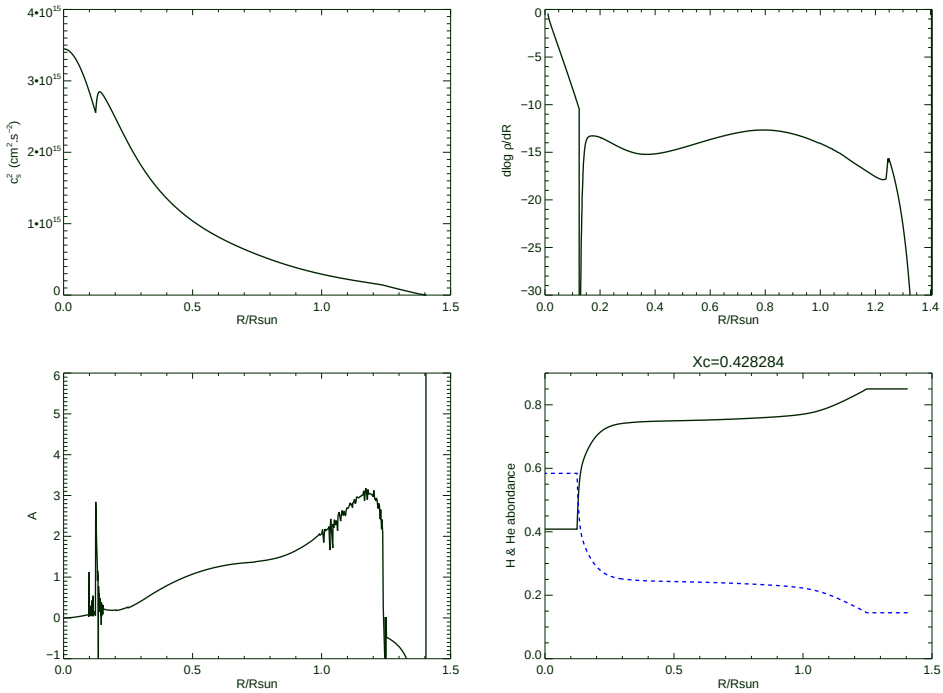


FIGURE 6.5 – **HD49933.** Vitesse du son, dérivée première de la densité, terme A et abondances en Hydrogène et hélium en fonction du rayon de l'étoile dans le scénario avec overshoot et avec diffusion. Un petit coeur convectif est présent. Des problèmes numériques apparaissent sur le terme A et la détermination des positions des zone convectives n'est pas aisée sur ce terme. En se basant sur le $dlog\rho/dR$, elle ne semble toutefois ne pas avoir significativement bougé. L'étoile est toujours à mi-vie environ. Notons aussi l'effet diffusif en surface qui abaisse fortement l'abondance en hélium.

compte de plus d'information qu'une simple moyenne. On notera toutefois qu'à 2σ, toutes les oscillations visibles ne sont pas significatives au regard de la moyenne (*i.e.* une droite horizontale passe par tous les points si les erreurs sont prises égales à 2σ).

Tous les ajustements que nous avons effectués se sont fait à physique identique. Celle-ci est exactement la même que celle utilisée pour HD49933, à 2 exceptions près : nous n'avons pas exploré de modèles avec diffusion et nous avons fixé le coefficient de mélange convectif à 0.7 (correspondant approximativement à la valeur solaire). Les observables non sismiques, issues de publications antérieures, on déjà été présentées dans le Chap.4.3.2 et sont résumées dans la Table 6.5, au même titre que les contraintes sismiques. Lors des ajustements, nous avons rencontré des difficultés similaires à celles décrites dans le cas de HD49933 : les résultats de la minimisation sont hautement sensibles aux paramètres initiaux, ce qui signifie qu'il existe un nombre important de minima locaux ou de fortes interdépendances entre paramètres. On notera par exemple que la grande séparation moyenne peut être reproduite très facilement par une large famille de modèles différant par leur masse et age : en augmentant l'un et en diminuant l'autre, la grande séparation moyenne est inchangée (anti-correlation).

Nous avons recherché quels sont les modèles reproduisant au mieux les observables pour une grande variété de conditions initiales. Au final, deux familles de solutions se dégagent.

1. Des solutions pour lesquelles la luminosité et la température effective est assez bien repro-

duite. Cette famille est caractérisée par une masse autour de $1.35M_\odot$ et un âge autour de 2.6 Milliards d'années.

2. Inversement, des solutions pour lesquelles on obtient une luminosité et température trop élevée, mais reproduisant un peu mieux les contraintes sismiques. Cette famille est caractérisée par une masse autour de $1.40M_\odot$ et un âge autour de 1.8 Millards d'années.

Bien que la valeur moyenne de la grande séparation soit correctement reproduite, aucune de ces deux solutions ne la reproduit aussi bien que pour HD49933 (phase et amplitude). Sur les Fig.6.7 et Fig.6.8, on a représenté l'accord entre les contraintes sismiques utilisées et les résultats du meilleur modèle, respectivement pour le cas 1 et 2 (appelés dans ce qui suit modèle 1 et modèle 2). Il semblerait que le modèle plus massif reproduit un peu mieux la petite séparation. Dans les deux cas, la quantité d'overshoot nécessaire est raisonnable ($\alpha_{ov} = 0.2$) et similaire. On trouvera en détail les résultats de la minimisation dans la Table 6.4.

Comme la grande séparation est mal reproduite dans chacun des modèles calculés, il est probable que des écarts importants existent entre les structures internes que nous montrons Fig.6.9 et Fig.6.10 et la réalité. Ceux-ci révèlent que l'étoile est probablement constituée d'un petit coeur convectif et d'une fine enveloppe externe. L'abondance en hydrogène au centre ne fait pas d'elle une étoile quittant la séquence principale comme nous l'avons suggéré précédemment. Cependant, ces modèles sont plutôt indicatifs que démonstratifs : tant que nous ne serons pas capable de reproduire précisément les variations de la grande séparation, nous ne pourrons prétendre détenir une structure interne représentative de HD181420. Précisons que la réduction des données effectuée (à savoir, une fonction sinusoïdale + constante pour la grande séparation) peut être en cause dans notre incapacité à retrouver le meilleur compromis entre la grande séparation observée et celle modélisée.

Les χ^2 obtenus sont en faveur du modèle 1, essentiellement car la luminosité, la température, l'abondance initiale du modèle $(Z/X)_0$ sont à plus de 1σ de la contrainte pour le modèle 2. Mais en terme sismique les 2 modèles sont relativement similaires.

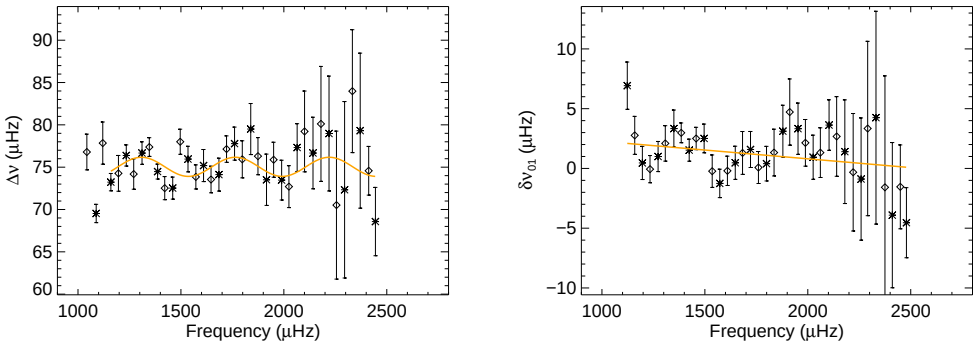


FIGURE 6.6 – **HD181420.** Ajustement préliminaire (étape de réduction de données) servant de base lors de la recherche du meilleur modèle d'intérieur stellaire. On remarque que certains points sont mal décrits par le terme sinusoïdal pour la grande séparation. De manière similaire à HD49933, la petite séparation semble osciller.

Au final, bien que nous devions rester prudents, nous pouvons dire que HD181420 est probablement dans le dernier tiers de sa séquence principale. Elle se caractériserait par de petites zones convectives interne et externe dont l'extension totale ($\approx 20\%$ du rayon) est similaire à celle de HD49933.

	Modèle 1	Modèle 2
$Age\ (Mans)$	2.642 ± 0.160	1.775 ± 0.140
$Masse\ (M_\odot)$	1.36 ± 0.11	1.40 ± 0.12
$R/R_\odot$	1.64	1.64
$(Z/X)_0$	0.0244 ± 0.0031	0.0259 ± 0.0024
Y_0	0.256 ± 0.04	0.274 ± 0.05
Xc	0.36	0.35
α_{CGM}	0.7	0.7
α_{ov}	0.21 ± 0.06	0.20 ± 0.07
$L/L_\odot$	0.645	0.69
T_{eff}	6543	6719
Epaiss. $Z_C^{inf}(\%R_{star})$	8.5%	7.0%
Epaiss. $Z_C^{sup}(\%R_{star})$	14.0%	10.6%
χ^2	8.3	51.9

TABLE 6.4 – **HD181420.** Principaux résultats après minimisation pour les 2 groupes de solutions obtenues. Les deux avant dernières lignes indiquent l'extension des zones convectives en fraction de rayon stellaire. Les deux modèles sont relativement similaires aux incertitudes près, excepté pour l'âge.

	$Age\ (Mans)$	$Masse\ (M_\odot)$	$(Z/X)_0$
a priori	$U[1000, 5000]$	$U[1.1, 1.6]$	$U[0.006 - 0.04]$
	Y_0	α_{CGM}	α_{ov}
a priori	$U[0.245, 0.350]$	$U[0.4, 1.0]$	$U[0.0, 0.6]$

TABLE 6.5 – **HD181420.** Plages tolérées lors de la recherche du meilleur modèle stellaire.

$T_{eff}\ (K)$	$log(L/L_\odot)$	$(Z/X)_S$	$C_0\ (\mu Hz)$	$A_0\ (\mu Hz)$
6580 ± 105	0.63 ± 0.03	0.0245 ± 0.004	75.1 ± 0.5	1.2 ± 0.3
$\omega\ (rad.ks)$	ϕ	$a\ (\mu Hz)$	b	
13.8 ± 1.6	-3.9 ± 0.9	3.8 ± 1.2	-1.5 ± 0.7	

TABLE 6.6 – **HD181420.** Résumé des contraintes observationnelles appliquées lors de la recherche du meilleur modèle stellaire. Toutes les contraintes sont gaussiennes car nous effectuons une minimisation du χ^2.

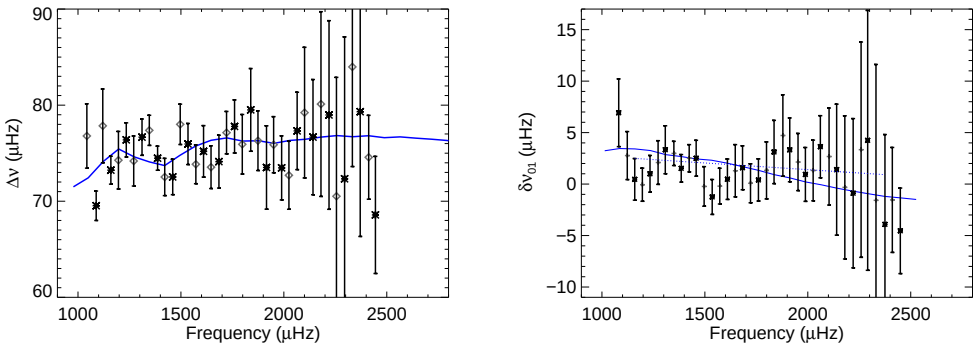


FIGURE 6.7 – **HD181420.** Meilleur ajustement trouvé (modèle 1). Ce modèle ne reproduit qu'assez mal les variations de la grande séparation, notamment la phase. La petite séparation est assez bien reproduite en moyenne.

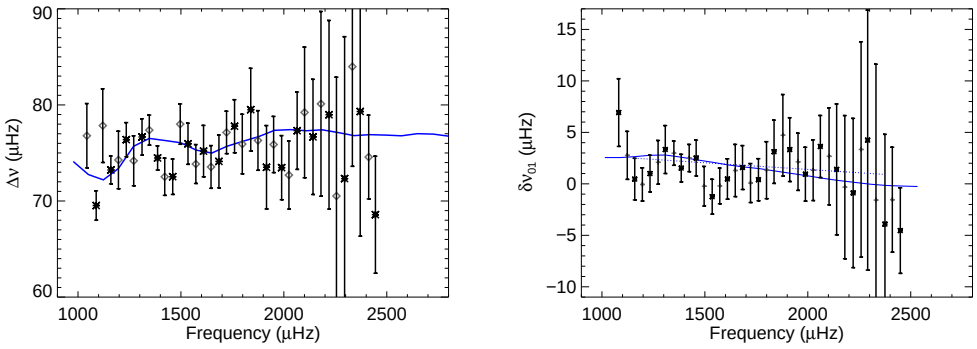


FIGURE 6.8 – **HD181420.** 2ème meilleur ajustement (modèle 2). Bien que ce soit le meilleur modèle obtenu, il ne reproduit qu'assez mal les variations de la grande séparation aux alentours de 1500 μHz notamment la période. La petite séparation est quant à elle bien reproduite en moyenne. Ce modèle reproduit aussi assez mal la température et la luminosité de l'étoile.

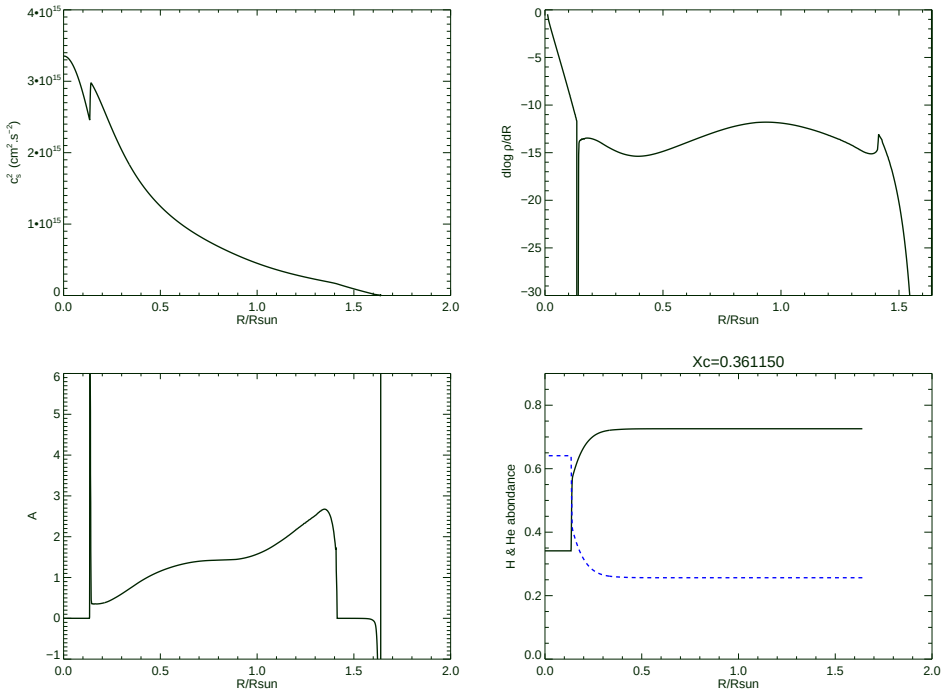


FIGURE 6.9 – **HD181420.** Vitesse du son, dérivée première de la densité, terme A et abondances en hydrogène et hélium en fonction du rayon de l'étoile pour la solution 1. Un petit coeur convectif et une petite enveloppe convective sont présents. Cette étoile est dans la séquence principale au vu de son abondance en hydrogène central.

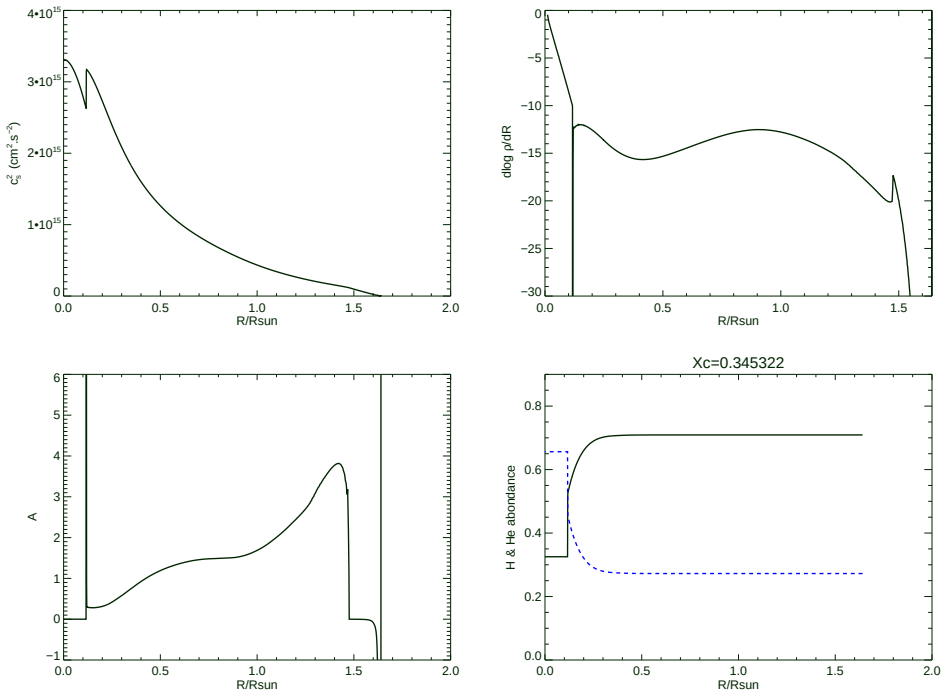


FIGURE 6.10 – **HD181420.** Vitesse du son, dérivée première de la densité, terme A et abondances en hydrogène et hélium en fonction du rayon de l'étoile dans la solution 2. La structure interne n'est pas radicalement modifiée comparativement au modèle précédent, plus âgé mais moins massif.

Chapitre 7

Conclusion et perspectives

Sommaire

7.1 Quelques mots sur les différences entre fréquentistes et bayésiens . 161
7.2 L'extraction de modes de pulsation 162
7.3 Des modes mixtes à foison 163
7.4 Interprétation : une critique de la méthode 165
7.5 Perspectives à moyen et long terme 166

7.1 Quelques mots sur les différences entre fréquentistes et bayésiens

Nous avons présenté et utilisé tout au long de ce manuscrit un ensemble d'outils statistiques pour l'analyse et l'interprétation d'observations. Deux interprétations des statistiques s'opposent. Tout d'abord, la vision fréquentiste présuppose que le monde est constitué de deux entités clairement séparables : des composantes aléatoires dont le comportement est imprévisible par essence mais suit des lois statistiques ; et des valeurs "vraies". Ces deux réalités se superposeraient pour construire le monde que nous mesurons avec nos instruments. Dans la vision bayésienne, cette dualité n'existe pas. Par contre elle en introduit une autre : ou bien nous avons l'information nécessaire pour juger d'une situation ou bien nous ne l'avons pas. Le propre du bayésien est donc de définir la probabilité comme une mesure de la véracité d'une hypothèse, ce que recherche justement un scientifique.

Une autre différence fondamentale réside dans le fait que l'approche bayésienne autorise l'utilisation d'information extérieure à l'objet de l'étude. On peut dans ce cadre construire à partir d'une information *a priori* (avant confrontation avec les faits) dont la nature peut être soit objective, soit subjective, une information *a posteriori*, c'est à dire après la confrontation avec les faits. En conséquence, des observateurs différents peuvent aboutir à des conclusions différentes et c'est justement ce que lui reproche les fréquentistes. Mais un processus scientifique n'est rien d'autre qu'un processus de hiérarchisation et de révision continuelle de connaissances se basant sur des *a priori* (des lois, des mesures antérieures ou même des hypothèses). L'approche bayésienne est donc un cadre formel (statistique) adapté aux sciences. Par opposition, la vision fréquentiste qui revendique l'objectivité du résultat obtenu par inférence n'est que difficilement applicable en sciences, car elle nécessite d'explorer toutes les éventualités possibles : ce n'est donc qu'une fiction théorique. En conclusion, en tout scientifique se cache un bayésien. Le cadre bayésien n'est rien d'autre qu'une manière de formaliser de l'information.

7.2 L'extraction de modes de pulsation

L'analyse des premières données du satellite CoRoT a révélé qu'il était bien difficile d'extraire une information fiable sur les étoiles plus chaudes que le soleil (type spectral F). Les techniques classiques d'inférence jusque là utilisées avec succès dans le cas solaire ont révélées leurs limites et il a fallu repenser intégralement le problème de l'ajustement des modes stellaires. C'est principalement ce constat qui à l'origine a motivé l'utilisation d'une approche bayésienne et de technique d'échantillonnage par des simulations Monte-Carlo avec des chaines de Markov (MCMC). Comme nous l'avons montré, l'échantillonnage permet d'extraire dans son intégralité la fonction de densité de probabilité, de mettre en évidence les corrélations entre paramètres et de calculer la probabilité *a posteriori* d'un modèle. Cela en fait un outil de diagnostic très puissant. En contre partie, le processus d'échantillonnage est environ 50 à 100 fois plus long que les méthodes par recherche de maximum.

Appliqué à HD49933 et HD181420, parmi les premières étoiles observées par le programme astérosismique de CoRoT, les outils développés durant la thèse ont montré leur efficacité. Par exemple dans le cas de HD49933, lors de l'analyse des données initiales (IRa-01), là où la méthode par maximum de vraisemblance privilégiait significativement la mauvaise identification des degrés des modes, un algorithme MCMC aboutit à un résultat bien plus contrasté : le processus décisionnel au sens bayésien n'est pas capable de trancher. Par contre, à partir du moment où nous avons disposé d'une information plus substantielle sur l'étoile (LRa-01), les méthodes par maximum de vraisemblance et l'approche bayésienne couplée à un algorithme d'échantillonnage ont fourni des résultats identiques. Concernant HD181420, son analyse nous a permis d'aboutir à une conclusion quant à l'identification des modes, là où Barban et al. (2009) ne put trancher. L'approche bayésienne couplée à un algorithme d'échantillonnage est donc un outil difficilement contournable lorsque nous sommes confrontés à des conditions difficiles d'analyse (fonction de vraisemblance avec multiples maxima) : le calcul bayésien de probabilité est plus robuste que le simple test de maximum de vraisemblance.

Toutes les approches par échantillonnage ont le souci d'être extrêmement couteuses en temps de calcul et il est donc difficilement envisageable de les utiliser sur un large nombre d'étoiles à l'heure actuelle. C'est pourquoi nous avons cherché à établir une méthode d'analyse du signal automatique qui soit rapide et relativement robuste pour déterminer les fréquences de pulsation des modes. Pour cela, on a mis en place une méthode originale conjuguant l'approche MCMC et MAP.

On a alors montré à titre de validation de la méthode, qu'il était possible de suivre les variations cycliques de l'activité solaire sur presque tous les paramètres des modes, bien que la résolution spectrale soit faible. Une nouvelle conséquence de l'activité solaire a même pu être mise en évidence : une modulation de la hauteur relative du $l = 2$. Cette modulation n'est pas visible pour les degrés $l = 1$ pourtant plus intenses. Elle pourrait être due à une plus forte sensibilité du $l = 2$ au champ magnétique des taches solaires, distribuées préférentiellement à basse latitude.

Ce genre d'outil d'analyse pourra donc être appliqué aux étoiles Kepler, lorsque nous bénéficierons d'observations longues (3 ans) voir très longues (si la mission est prolongée). Bien qu'évidemment, avec le Soleil, nous bénéficions d'un rapport hauteur des modes sur bruit très élevé, nous pensons que la méthode est applicable à d'autres étoiles. En effet, dans la mission Kepler, plusieurs astres ont des niveaux hauteurs sur bruit seulement deux fois inférieurs à ceux constatés dans les données de l'instrument VIRGO.

L'application de notre méthode à des spectres synthétiques d'étoiles a aussi montré qu'il était possible pour près de 78% des étoiles ayant un rapport signal à bruit significatif et faisant partie de la séquence principale, d'établir l'identification correctement et automatiquement. Malheu-

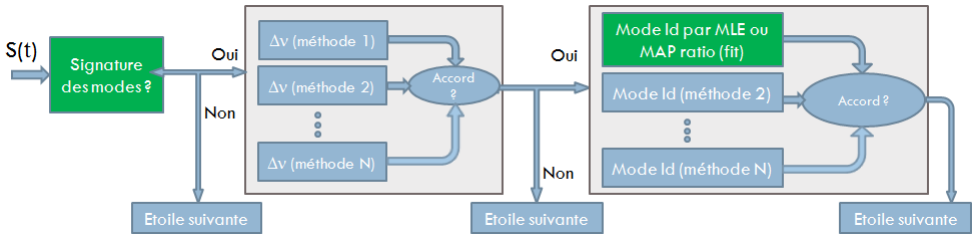


FIGURE 7.1 – Vue schématique simplifiée des étapes fondamentales dans ce que devrait être une procédure pleinement automatique d'ajustement des modes. En vert sont représentées les parties décrites (et implémentées) dans le Chap.5.

resement en l'état actuel, les hypothèses formulées ne permettent pas d'ajuster les modes de pulsation des étoiles évoluées, essentiellement à cause de la présence de modes mixtes.

Le type d'approche présenté ici est bien évidement perfectible, mais il ouvre des perspectives concernant le problème de l'ajustement d'une grande quantité d'étoiles en un temps raisonnable [1]. Par exemple, obtenir une valeur précise de la grande séparation est une donnée critique pour que la convergence se fasse correctement dans le cas d'une procédure automatique. Nous pensons donc que la détermination de cet *a priori* doit se faire sur la base multiples méthodes dont les résultats doivent être mutuellement vérifiés (procédure de *cross checking*). Il en est de même pour l'identification. On pourrait ainsi établir une classification en fonction du degré de confiance sur le résultat obtenu. Sur la Fig.7.1, nous avons représenté schématiquement les étapes principales de ce que pourrait être, de notre point de vue, une procédure pleinement automatique d'ajustement des modes p. On retrouvera en vert les parties que j'ai implémenté.

7.3 Des modes mixtes à foison

Au delà de l'analyse des données CoRoT, je suis activement impliqué dans l'analyse des étoiles du satellite Kepler. Plusieurs étoiles de ces données fort intéressantes ont déjà été analysées avec l'un ou l'autre des outils précédemment présentés. Nous n'avons pu malheureusement nous attarder dessus car la politique de publication des données Kepler est relativement stricte. Parmi les milliers d'étoiles à disposition, il en existe un nombre très important (plus de la moitié) avec des modes mixtes. Les étoiles les plus âgées ont des pulsations d'amplitudes plus fortes que leur cadettes et c'est probablement à cause de cela qu'elles sont plus représentées. Ainsi, une grande part des étoiles observées présentent des modes mixtes. Bien que très intéressant à interpréter, ils sont extrêmement difficiles à extraire du spectre de puissance. On a par exemple représenté sur la Fig.7.2 le diagramme-échelle obtenu pour une des étoiles Kepler observée durant 229 jours. Cette étoile est assez bien représentative de l'ensemble des étoiles que nous avons à analyser avec Kepler. On remarque un grand nombre de modes mixtes pour le degré $l = 1$ dont certains peuvent se trouver très proches du $l = 0$. Cette proximité complique la tâche d'extraction car il devient aisé de confondre le degré des modes. La durée d'observation extrêmement longue permet d'obtenir une détermination très précise des fréquences d'oscillations.

Cette nature mixte des modes peut aussi rendre difficile la recherche de modèle de structure interne optimale. Par exemple, la grande et la petite séparation ne sont plus des critères pertinents car la loi asymptotique n'est plus une approximation valide. Une comparaison fréquence

1. Il n'a fallut qu'environ 5 jours de calculs pour ajuster les 177 étoiles sur une machine standard.

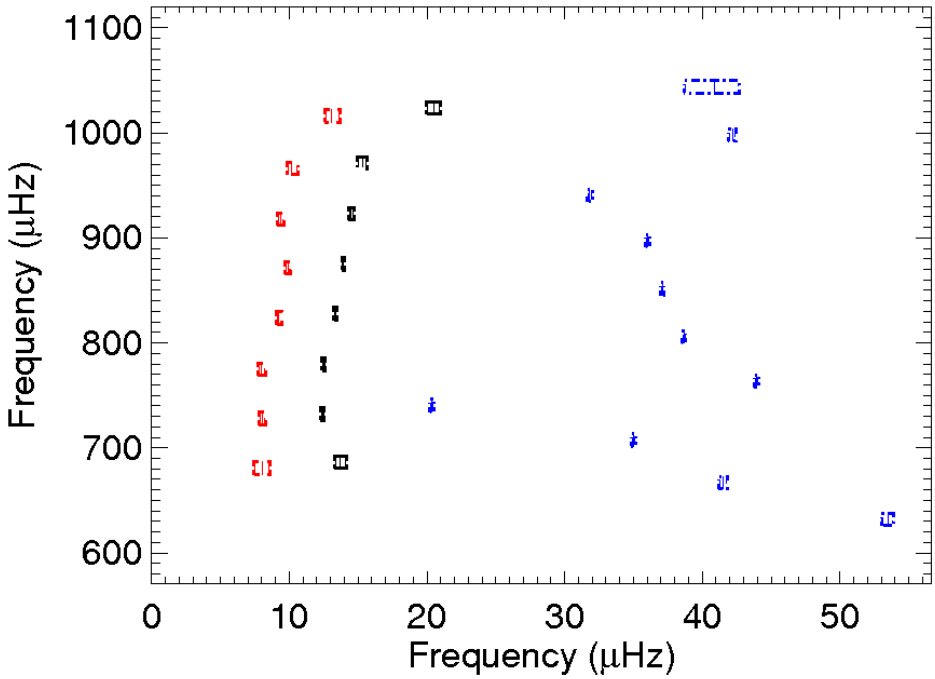


FIGURE 7.2 – Diagramme échelle d'une des étoiles observée par Kepler. On remarquera la forte dispersion des degrés $l = 1$ (en bleu) qui sont en grand nombre mixtes. Les degrés $l = 0$ et $l = 2$ sont en noir et en rouge. Les intervalles de confiance à 1σ sont symbolisés par une boite. On remarque l'extrême précision sur la détermination des fréquences : l'erreur à 1σ est de l'ordre 0.1 à 0.4 μHz.

à fréquence est indispensable. Compte tenu des contraintes extrêmement fortes fournies par les modes mixtes, il devient aussi *a priori* très difficile de trouver par une procédure de minimisation un modèle adéquat reproduisant toutes les fréquences observées. Des grilles de modèles peuvent faciliter la tâche. Quoiqu'il en soit, il y a un travail très intéressant à effectuer pour déterminer à l'avenir quelle(s) méthode(s) utiliser pour déterminer les modèles de structure interne reproduisant au mieux les observables des étoiles ayant des modes mixtes.

7.4 Interprétation : une critique de la méthode

Dans le dernier chapitre du présent manuscrit, nous avons montré qu'avec une approche relativement simple nous pouvions déterminer des modèles de structure interne d'étoiles satisfaisant plus ou moins des contraintes sismiques, lorsque la loi asymptotique est valide. Nous avons cependant insisté sur le fait que cette approche, mettant en oeuvre un algorithme de Levenberg-Marquardt (LM) est sensible aux conditions initiales. Les limitations dues à la modélisation, notamment celles concernant le calcul des fréquences théoriques d'oscillation (effets de surface mal pris en compte), contribuent aussi à rendre la tâche d'interprétation quelque peu difficile (ou tout du moins difficilement automatisable). Nous avons aussi brièvement dit que la détermination conjointes de certaines variables stellaires (tel que la masse, l'âge, et le coefficient de convection) lors de la minimisation peut aussi être source de problème. En effet, il existe de très fortes corrélations entre ces variables formant des "vallées" de stabilité dans l'espace des paramètres. En plus des minima locaux, des régions d'espace très "plates" peuvent mettre en échec l'algorithme LM : ce dernier peut alors converger tout le long de la vallée et ne jamais trouver le véritable maximum dans l'espace des paramètres. Il est évident que de larges progrès peuvent être effectués concernant la recherche de modèles optimums.

Si nous ne souhaitons pas tomber dans les mêmes pièges que les méthodes MLE appliquées au spectre de puissance de HD49933, il est selon moi impératif d'établir des analyses plus robustes et intensives du signal. De mon point de vue, ceci passe nécessairement par l'utilisation d'outils statistiques plus performants (comme les MCMC mais pas nécessairement. Les réseaux de neurones offrent une bonne solution alternative ou même tout simplement la constitution de grilles de modèles). Le coût en temps de calcul certes s'accroit (la puissance de calcul n'est qu'un problème de moyens informatiques), mais je pense que c'est une étape indispensable pour construire des modèles véritablement cohérents avec les observables à disposition et obtenir des résultats robustes. J'ajouterais que la recherche du véritable maximum de probabilité avec un algorithme LM est longue et fastidieuse, sans que cela ne garantisse aucunement la pertinence de la solution trouvée. Au final, l'effort mis à rechercher le maximum a de fortes chances d'être plus important que celui qui aurait été fait pour échantillonner intégralement les fonctions de densité de probabilités des paramètres via une méthode MCMC : le temps perdu à rechercher la meilleure solution par un algorithme LM peut être dépensé en temps d'échantillonnage.

De plus, dans le cas présent, j'ai dû supposer toutes les erreurs gaussiennes. Or ce n'est pas systématiquement le cas pour les fréquences : certaines fréquences sont multimodales et réduire l'information (comme je l'ai fait ici même) à des statistiques gaussiennes est une hypothèse forte difficilement soutenable. L'utilisation de méthodes d'échantillonnage MCMC permettraient d'extraire une information bien plus riche et d'utiliser intégralement l'information obtenue lors des étapes précédentes (extraction des fréquences des modes) sans émettre aucune hypothèse réductrice sur la nature de la statistique des contraintes observationnelles. Une autre conséquence de l'échantillonnage serait de permettre une véritable estimation des corrélations et erreurs dans l'hyper-espace des paramètres sur toutes les grandeurs physiques d'intérêt et pas seulement celles ajustées (comme la position de la zone convective, le rayon ou l'abondance en hydrogène au coeur

par exemple). Contrairement aux méthodes par maximum (MLE/MAP) qui ne fournissent que le modèle optimal, l'échantillonnage a l'avantage de fournir une famille de modèles cohérents avec les observables : il est donc possible de reconstruire la distribution *a posteriori* (donc de calculer les erreurs) de n'importe quelle variable, implicitement ajustée ou non.

On peut estimer ce qui peut être fait en ce sens par quelques calculs simples. En supposant que nous fassions varier un maximum de 5 paramètres, il faut compter environ 100 000 à 200 000 échantillons pour obtenir des distributions de probabilité échantillonnées de manière suffisante. Si on se fixe une limite en temps de calcul de 30 jours, on aboutit [2] à un temps de calcul maximum par modèle de 13 à 26 secondes. Les MCMC étant facilement parallélisables, si on dédie une machine octo-processeur (machines peu coûteuses de nos jours) actuelle, on peut multiplier par environ 8 ce temps, soit un temps maximum autorisé par modèle de 1 min 42 sec à 3 min 27 sec. Les modèles calculés sans diffusion présentés ici nécessitent environ 6 minutes, ce qui est en dehors de la fourchette. Cependant, en réduisant la précision des modèles (notamment en augmentant les pas de temps) il est déjà possible de calculer des modèles relativement précis en 2 à 3 minutes avec CESAM2k [3]. De plus, en tenant compte de l'accroissement exponentiel de la puissance des machines [4], ce type de calcul pourra être mené avec une précision croissante dans les prochaines années. Enfin, ajoutons qu'un cluster (ou un supercalculateur) pourrait faire de même en quelques jours. Il est toutefois difficilement envisageable de calculer des modèles dont la physique est relativement complexe. Par exemple, avec de la diffusion, le temps de calcul par modèle peut être de l'ordre de l'heure.

Dans un avenir proche, je pense développer et appliquer de tels moyens statistiques à la modélisation. Cela permettra de mieux cerner les dépendances réciproques entre paramètres. Il sera alors éventuellement envisageable d'établir une re-paramétrisation, à l'image de celle que nous avons effectuée dans le Chap.5.3.2 (introduction explicite de la corrélation entre les termes du polynôme d'ordre 2 décrivant les fréquences des modes). Une telle re-paramétrisation pourrait aider un algorithme LM à converger correctement vers le véritable maximum de probabilité. A titre d'exemple, les paramètres âge et masse de l'étoile sont à l'évidence corrélés. De fait, je n'ajuste probablement pas toujours les paramètres les plus pertinents pour aboutir à une solution unique en l'état actuel.

7.5 Perspectives à moyen et long terme

Globalement, ma thèse m'a permis d'entrevoir quels seront les grandes difficultés auxquelles nous serons confrontés. Deux catégories de difficultés sont identifiables.

Tout d'abord d'ordre scientifique. Les codes de modélisation stellaires permettent d'ores et déjà de décrire assez bien la structure interne des étoiles lorsque les phénomènes en jeu ne sont pas extrêmes. Or, de par la grande diversité des objets célestes, il existe une pléthore d'objets pour lesquelles ce n'est plus le cas. Pour ne citer qu'un exemple : une rotation brise la symétrie sphérique. Lorsque cette rotation est rapide, la théorie des perturbations n'est alors plus forcément applicable et un calcul unidimensionnel n'est plus possible. Enfin, dans la diversité des astres qui nous entourent il existe très probablement une large gamme de phénomènes physiques qui sont à découvrir et à comprendre.

2. Avec un CPU Intel Xeon cadencé à 3 Ghz.
3. On peut aussi envisager d'utiliser un autre code d'évolution, éventuellement plus rapide.
4. La conjecture de Moore (ou loi de Moore), stipule que le nombre de transistor d'un microprocesseur double tous les deux ans. Bien que subissant un léger ralentissement ces dernières années, cette loi est vérifiée depuis le début des années 1970.

Mettre en évidence ces phénomènes physiques nécessite d'avoir des mesures de qualité et dont l'interprétation est la moins ambiguë possible. Ces objectifs ne peuvent être atteints qu'avec des progrès instrumentaux et techniques en générale. Vu la complexité des phénomènes physiques en jeu, la qualité des données et leur analyse est en effet un obstacle. Une question qui se pose couramment est : comment discriminer un signal dû à un phénomène physique et une perturbation ou un bruit ?

Plus particulièrement, il est très difficile d'extraire des données une information la plus objective possible à partir d'informations *a priori* parcellaires voir incorrectes. Ces situations sont courantes en science car le travail de recherche se situe par définition à la frontière de nos connaissances. Ce sont donc essentiellement les processus décisionnels qui sont problématiques et sur lesquelles nous devons porter une grande attention à l'avenir. A cette difficulté s'ajoute l'incommensurable nombre d'objets observés et la nature hétéroclite des phénomènes qui en découle. Les capacités humaines étant limitées, nous serons très vite incapable d'analyser ou d'interpréter correctement la masse de données qui s'accumulera (nous sommes d'ores et déjà mis en difficulté avec Kepler) : les capacités humaines de décision et de jugement sont certes probablement les meilleures actuellement, mais bientôt nous ne pourrons gérer l'aflux d'information important auxquels nous faisons face. Par exemple, PLATO devrait observer plus de 500 000 étoiles sur 3 ans si il est lancé aux alentours de 2018, ce qui est 5 fois plus que Kepler. Si nous souhaitons analyser au mieux toutes ces données il nous faudra impérativement nous reposer sur des processus automatiques de décision et de classification : y a t il des modes g ? y a t il des modes p ? des modes mixtes ? si oui, quelle est l'identification de ces modes ? quelle est la vitesse de rotation de surface ? on ne peut répondre à toutes ces questions et à bien d'autres sans automate.

Une réflexion en profondeur doit donc être menée sur ce point, en s'inspirant et en collaborant avec d'autres communautés scientifiques déjà confrontées à des problèmes similaires (les roboticiens ou les physiciens des particules par exemple). Ce sera un travail de longue haleine, dans lequel il faudra se lancer dans les années à venir.

Bibliographie

Aerts, C., Christensen-Dalsgaard, J., & Kurtz, D. 2010, Springer

Aigrain, S., Favata, F., & Gilmore, G. 2004, *Astron. Astrophys.*, 414, 1139

Aitken, A. & Silverstone, H. 1942, Proc. Roy. Soc. Edin., A, 62, 369

Alexander, D. R. & Ferguson, J. W. 1994, *Astrophys. J.*, 437, 879

Anderson, E. R., Duvall, Jr., T. L., & Jefferies, S. M. 1990, *Astrophys. J.*, 364, 699

Ando, H. & Osaki, Y. 1975, *Pub. Astron. Soc. Japan*, 27, 581

Andrieu, C. & Robert, C. P. 2001, Controlled MCMC for Optimal Sampling, Tech. rep.

Appourchaux, T. 2003, *Astron. Astrophys.*, 412, 903

Appourchaux, T., Gizon, L., & Rabello-Soares, M.-C. 1998, *Astron. Astrophys. Suppl.*, 132, 107

Appourchaux, T., Michel, E., Auvergne, M., et al. 2008, *Astron. Astrophys.*, 488, 705

Asplund, M., Grevesse, N., & Sauval, A. J. 2005, in Astronomical Society of the Pacific Conference Series, Vol. 336, Cosmic Abundances as Records of Stellar Evolution and Nucleosynthesis, ed. T. G. Barnes III & F. N. Bash, 25

Atchadé, Y. F. 2006, Meth. Comp. In Applied Probab., 8, 235

Badnell, N. R., Bautista, M. A., Butler, K., et al. 2005, *Mon. Not. Roy. Astron. Soc.*, 360, 458

Baglin, A., Auvergne, M., Barge, P., et al. 2006, ESA Special Publication, 1306, 33

Ballot, J. 2010, Astronomichen Nachrichten

Ballot, J., Appourchaux, T., Toutain, T., & Guittet, M. 2008, *Astron. Astrophys.*, 486, 867

Barban, C., Deheuvels, S., Baudin, F., et al. 2009, *Astron. Astrophys.*, 506, 51

Benomar, O. 2008, Communications in Asteroseismology, 157, 98

Benomar, O., Appourchaux, T., & Baudin, F. 2009a, *Astron. Astrophys.*, 506, 15

Benomar, O., Baudin, F., Campante, T. L., et al. 2009b, *Astron. Astrophys.*, 507, L13

Bevington, P. R. & Robinson, D. K. 2003, Data reduction and error analysis for the physical sciences, ed. Bevington, P. R. & Robinson, D. K.

Böhm-Vitense, E. 1958, Zeitschrift fur Astrophysik, 46, 108

Borucki, W. J., Koch, D. G., Lissauer, J., et al. 2007, in Astronomical Society of the Pacific Conference Series, Vol. 366, Transiting Extrapolar Planets Workshop, ed. C. Afonso, D. Weldrake, & T. Henning, 309

Brewer, B. J., Bedding, T. R., Kjeldsen, H., & Stello, D. 2007, *Astrophys. J.*, 654, 551

Brooks, S. & Roberts, G. 1998, Statistics and Computing, 8, 319

Bruntt, H. 2009, A&A, 506, 235

Bruntt, H., De Cat, P., & Aerts, C. 2008, *Astron. Astrophys.*, 478, 487

Burgers, J. M. 1969, Flow Equations for Composite Gases, ed. Burgers, J. M.

Canuto, V. M., Goldman, I., & Mazzitelli, I. 1996, ApJ, 473, 550

Catala, C., Arentoft, T., Fridlund, M., et al. 2010, in Astronomical Society of the Pacific Conference Series, Vol. 430, Astronomical Society of the Pacific Conference Series, ed. V. Coudé Du Foresto, D. M. Gelino, & I. Ribas, 260

Chaplin, W. J., Appourchaux, T., Arentoft, T., & et al. 2008, *Astronomische Nachrichten*, 329, 549

Chaplin, W. J., Elsworth, Y., Isaak, G. R., Miller, B. A., & New, R. 2000, MNRAS, 313, 32

Christensen-Dalsgaard, J. 1997

Cowles, M. K. & Carlin, B. P. 1996, Journal of the American Statistical Association, 91, 883

Cowling, T. G. 1941, *Mon. Not. Roy. Astron. Soc.*, 101, 367

Cox, J. 1984, Gordon and Breach Science Publishers

Cramer, H. 1946, Pinceton Univ. Press

Cristian, P., Mihaela, P., & Cristian, V. C. 2004, Journal of Chemical Physics, 120, 4190

Daeppen, W., Mihalas, D., Hummer, D. G., & Mihalas, B. W. 1988, *Astrophys. J.*, 332, 261

De Ridder, J., Barban, C., Baudin, F., et al. 2009, *Nature*, 459, 398

Degroote, P., Aerts, C., Ollivier, M., et al. 2009, *Astron. Astrophys.*, 506, 471

Deheuvels, S., Bruntt, H., Michel, E., et al. 2010, *Astron. Astrophys.*, 515, A87

Deubner, F. & Gough, D. 1984, *ARA&A; KNUDSEN*, 22, 593

Deubner, F. L., Noyes, R. W., & Simon, G. W. 1975, *Astron. Astrophys.*, 371

Duvall, T. & Harvey, J. 1986, NATO ASI Series C : Mathematical and Physical Sciences, 169, 105

Earl, D. & Deem, M. W. 2005, Phys. Chem. Chem. Phys., 7, 3910

Eckhardt, R. 1987, Los Alamos Science Special Issue

Evans, J. W. & Michard, R. 1962, *Astrophys. J.*, 136, 493

Ferguson, J. W., Alexander, D. R., Allard, F., et al. 2005, *Astrophys. J.*, 623, 585

Formicola, A., Imbriani, G., Costantini, H., et al. 2004, Physics Letters B, 591, 61

Frohlich, C., Andersen, B. N., Appourchaux, T., & et al. 1997, *Solar Phys.*, 170, 1

Gamerman, D. 1997, Chapman & Hall/CRC

Garcia, R. A., Ballot, J., Eff-Darwich, A., et al. 2010, ArXiv e-prints

García, R. A., Régulo, C., Samadi, R., et al. 2009, *Astron. Astrophys.*, 506, 41

Gaulme, P., Schmider, F., & Gay, J. 2010, in Bulletin of the American Astronomical Society, Vol. 42, Bulletin of the American Astronomical Society, 1011

Gelman, ., Roberts, G., & W.R., G. 1996, Bayesian Statistics, 5, 599

Geyer, C. J. 1991, Computing Science Statistics Proc. 23th Symp. on the Interface, 5, 156

Gillon, M. & Magain, P. 2006, A&A, 448, 341

Gizon, L. & Solanki, S. K. 2003, *Astrophys. J.*, 589, 1009

Goldreich, P. & Keeley, D. A. 1977, *Astrophys. J.*, 212, 243

Gregory, P. C. 2005b, Bayesian Logical Data Analysis for the Physical Sciences : A Comparative Approach with *Mathematica* Support (Cambridge University Press)

Grevesse, N. & Noels, A. 1993, Cambridge University press, 15

Gruberbauer, M., Kallinger, T., & Weiss, W. W. 2009, arXiv :0811.3345v2

Harvey, J. 1985, ESA SP, 235, 199

Hastings, W. 1970, Biometrika, 57, 97

Hekker, S., Kallinger, T., Baudin, F., et al. 2009, *Astron. Astrophys.*, 506, 465

Hummer, D. G. & Mihalas, D. 1988, *Astrophys. J.*, 331, 794

Iben, Jr., I. & MacDonald, J. 1985, *Astrophys. J.*, 296, 540

Jeffrey, A. 1961, Cambridge University Press

Jennison, C. 1993, J. Roy. Statist. Soc. Series B, 55, 54

Katzgraber, H. G., Trebst, S., Huse, D. A., & Troyer, M. 2006, Journal of Statistical Mechanics : Theory and Experiment, 2006, P03018

Kendall, M. & Stuart, A. 1967, Charles Griffin and Company, 2

Kjeldsen, H. & Bedding, T. R. 1995, *Astron. Astrophys.*, 293, 87

Kjeldsen, H., Bedding, T. R., Arentoft, T., et al. 2008, *Astrophys. J.*, 682, 1370

Komm, R. W., Howe, R., & Hill, F. 2000, *Astrophys. J.*, 531, 1094

Leavitt, H. S. & Pickering, E. C. 1912, Harvard College Observatory Circular, 2

Léger, A., Rouan, D., Schneider, J., et al. 2009, *Astron. Astrophys.*, 506, 287

Leibacher, J. W. & Stein, R. F. 1971, *Astrophys. Lett.*, 7, 191

Leighton, R., Noyes, R. W., & Simon, G. W. 1962, Astrophys. J.

Lin, C.-Y., Hu, Chin-Kun, & Hansmann, H. 2003, PROTEINS : Structure, Function and Genetics, 52, 436

Marques, J. P., Monteiro, M. J. P. F. G., & Fernandes, J. M. 2008, *Astrophys. Spa. Sci.*, 316, 173

Mathur, S., García, R. A., Régulo, C., et al. 2010, *Astron. Astrophys.*, 511, A46

Metropolis, N., Rosenbluth, A., & Rosenbluth, M. 1953, J. Chem. Phys, 21, 188

Michel, E., Baglin, A., Auvergne, M., et al. 2008, Science, 322, 558

Mihalas, D., Dappen, W., & Hummer, D. G. 1988, *Astrophys. J.*, 331, 815

Monteiro, M. J. P. F. G., Christensen-Dalsgaard, J., & Thompson, M. J. 1994, *Astron. Astrophys.*, 283, 247

Morel, P. 1997, A&A, 124, 597

Morel, P. 2009, Utilisation et Description du Code d'évolution stellaire cesam2k (https ://svn.oca.eu/trac/planets/export/65/trunk/doc/Cesam2k/cesam2k.pdf)

Morel, P. & Lebreton, Y. 2008, ApSS, 316, 61

Mosser, B. & Appourchaux, T. 2009, *Astron. Astrophys.*, 508, 877

Mosser, B., Baudin, F., Lanza, A., et al. 2009a, *Astron. Astrophys.*, accepted

Mosser, B., Bouchy, F., Catala, C., et al. 2005, *Astron. Astrophys.*, 431, L13

Mosser, B., Michel, E., Appourchaux, T., et al. 2009b, *Astron. Astrophys.*, 506, 33

Powell, R. 2009, Atlas of the Universe : The Hertzsprung Russel diagram, http ://www.atlasoftheuniverse.com/hr.thml

Propp, J. G. & Wilson, D. B. 1996in , 223–252

Robbins, H. & Monro, S. 1951, Ann. Math. Stat., 22, 400

Robert, C. & Casella, G. 2008, Research note : Université Paris-Dauphine

Roberts, G. & Rosenthal, J. 2001, Statist. Sci., 16, 351

Roberts, G. O., Gelman, A., & Gilks, W. R. 1994, Tech Rep.

Roberts, G. O., Gelman, A., & Gilks, W. R. 1997, Ann. Appl. Prob., 7, 110

Rogers, F. J. & Iglesias, C. A. 1996, in Bulletin of the American Astronomical Society

Rogers, F. J. & Nayfonov, A. 2002, *Astrophys. J.*, 576, 1064

Roxburgh, I. W. 2009, *Astron. Astrophys.*, 493, 185

Saio, H. 1981, *Astrophys. J.*, 244, 299

Samadi, R., Georgobiani, D., Trampedach, R., et al. 2007, *Astron. Astrophys.*, 463, 297

Scargle, J. D. 1982, *Astrophys. J.*, 263, 835

Scuflaire, R., Montalbán, J., Théado, S., et al. 2008, *Astrophys. Spa. Sci.*, 316, 149

Solano, E., Catala, C., Garrido, R., et al. 2005, *Astronom. J.*, 129, 547

Stello, D., Chaplin, W. J., Basu, S., Elsworth, Y., & Bedding, T. R. 2009, *Mon. Not. Roy. Astron. Soc.*, 400, L80

Swendsen, R. H. & Wang, J.-S. 1986, Phys. Rev. Lett., 57, 2607

Tarrant, N. J. 2010

Tassoul, M. 1980, *Astrophys. J. Suppl. Ser.*, 43, 469

Thévenin, F., Bigot, L., Kervella, P., et al. 2006, Memorie della Societa Astronomica Italiana, 77, 411

Ulrich, R. K. 1970, *Astrophys. J.*, 162, 993

Unno, W., Osaki, Y., Ando, H., Saio, H., & Shibahashi, H. 1989, Springer

Vorontsov, S. 1988, in Advances in Hélio and Astéroseismology, ed. J. Christensen-Dalsgaard, & S. Fransden (Reidel), IAU, Symp., 123, 151

Walker, G., Matthews, J., Kuschnig, R., et al. 2003, *Pub. Astron. Soc. Pac.*, 115, 1023

Woodard, M. F. & Noyes, R. W. 1985, *Nature*, 318, 449

Zahn, J. 1991, *Astron. Astrophys.*, 252, 179

Annexe A

ANNEXE : bibliothèque de fonctions *a priori*

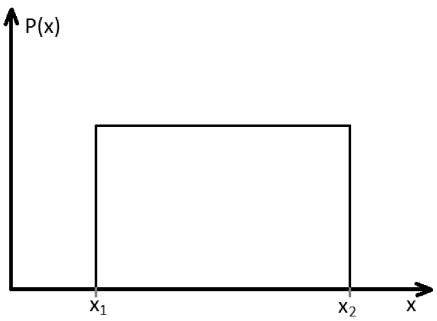
On distingue deux catégories d'*a priori*. Les *a priori* non informatif et les *a priori* informatif. Les premiers ont une variance σ^2 qui tend vers l'infini. Ce type d'*a priori* n'étant pas normalisable (l'intégrale diverge), en pratique ils ne sont qu'une fiction théorique : pour que l'intégrale sur le paramètre soit finie, l'espace est borné. Par exemple un *a priori* uniforme est considéré comme non informatif, mais ce n'est réellement le cas que si ces bornes tendent vers l'infini ($\pm\infty$). Ce type d'*a priori* est très utile lorsque nous n'avons qu'une connaissance très floue sur un paramètre. A contrario, un *a priori* informatif a une variance finie. Par exemple, un *a priori* gaussien est considéré comme informatif.

Dans ce qui suit nous présentons les quelques *a priori* que nous avons utilisés durant notre travail de recherche. Deux d'entre eux sont non informatifs : l'*a priori* (de Jeffrey) uniforme et l'*a priori* (de Jeffrey) uniforme dans l'espace logarithmique. Ce dernier est plus communément appelé *a priori* de Jeffrey.

***A priori* uniforme (U).** Cet *a priori* est uniforme sur la plage x_1 et x_2. On utilise la notation $U(x_1, x_2)$ pour le désigner. Il est de nature non-informative.

$$\forall\ x \in [x_1, x_2],\ P(x) = C$$
$$\forall\ x \in]-\infty, x_1[\cup]x_2, +\infty[= 0$$

avec C la constante de normalisation. $C = \frac{1}{x_2 - x_1}$.



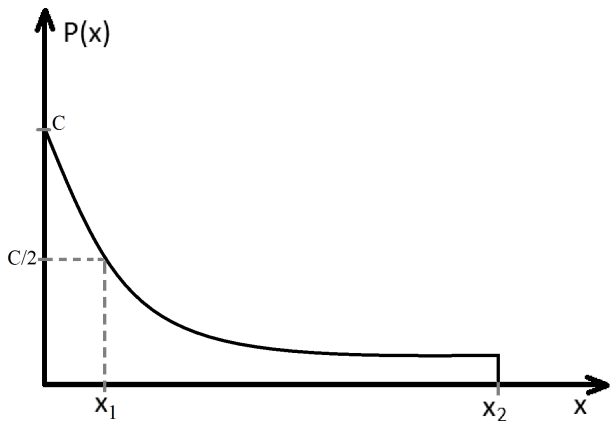
***A priori* de Jeffrey (J).** Cet *a priori* est uniforme dans l'espace logarithmique. Il est caractérisé par x_1 et x_2.

x_1 est défini par $P(x_1) = C/2$, C étant la constante de normalisation de l'*a priori*. C'est donc une grandeur caractéristique définissant la "vitesse" de décroissance de la fonction de probabilité de densité. En ajustant x_1 on peut plus ou moins fortement discriminer bruit et signal. Cet *a priori* donne plus de crédit au bruit *i.e.*, aux valeurs de faibles amplitudes. Si on dispose d'un ensemble de mesures x_i constituées que de bruit alors la probabilité *a posteriori* sera peu pénalisée par l'*a priori* ($x \approx 0$). A l'inverse, si un signal significatif est présent, la pénalisation est plus forte mais la probabilité *a posteriori* sera dominée par la vraisemblance car si $x >> x_1$ alors $P(x)$ est approximativement uniforme.

x_2 est une limite supérieure au delà de laquelle la fonction est nulle. x ne peut prendre que des valeurs positives, c'est pourquoi cet *a priori* ne peut être utilisé que sur des grandeurs intensives, par exemple les hauteurs ou les largeurs (par opposition au grandeurs de positions tel que la fréquence). On utilise la notation $J(x_1, x_2)$ pour le désigner. Il est de nature non-informative.

$$\forall\ x \in\ [0, x_2],\ P(x) = C\frac{1}{x+x_1}$$
$$\forall\ x \in\ ]-\infty, 0[\cup]x_2, +\infty[= 0$$

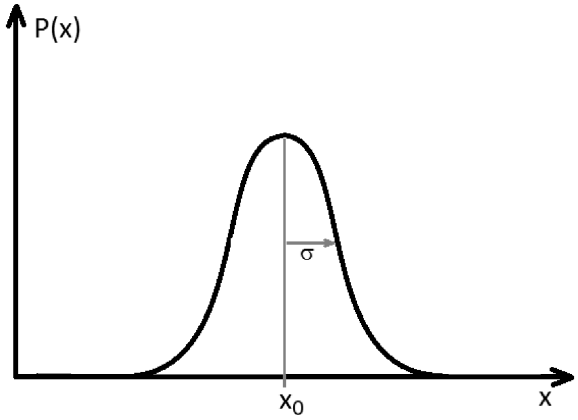
avec C la constante de normalisation. $C = ln\left(1 + \frac{x_2}{x_1}\right)^{-1}$.



***A priori* gaussien (N).** C'est une fonction gaussienne décrite par sa valeur centrale x_0 et son écart-type σ. On la note $N(x_0, \sigma)$ Il est de nature informative.

$$\forall\ x \in]-\infty, +\infty[= C\, exp(-\tfrac{1}{2}(x - x_0)^2/\sigma^2)$$

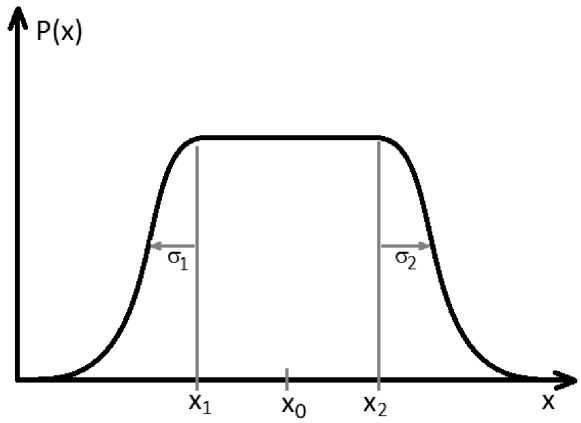
avec C la constante de normalisation. $C = \frac{1}{\sqrt{2\pi\sigma}}$.



***A priori* gaussien-uniforme-gaussien (GUG).** Cet *a priori* est uniforme sur la plage x_1 et x_2 et gaussien à ses deux extrémités. x_0 correspond à la valeur centrale défini comme $(x_1+x_2)/2$. σ_1 et σ_2 sont les largeurs caractéristiques (écart type) des ailes gaussiennes. On utilise la notation $GUG(x_1, x_2, \sigma_1, \sigma_2)$ pour le désigner. Il est de nature informative.

$$\forall\ x \in [x_1, x_2],\ P(x) = C$$
$$\forall\ x \in]-\infty, x_1[= C\, exp(-\tfrac{1}{2}(x-x_1)^2/\sigma_1^2)$$
$$\forall\ x \in]x_2, +\infty[= C\, exp(-\tfrac{1}{2}(x-x_2)^2/\sigma_2^2)$$

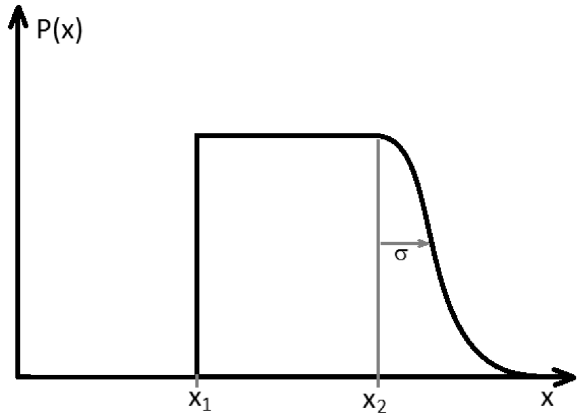
avec C la constante de normalisation. $C = \left((x_2 - x_1) + \frac{1}{2}\sqrt{2\pi}\sigma_1 + \frac{1}{2}\sqrt{2\pi}\sigma_2\right)^{-1}$.



***A priori* uniforme-gaussien (UG).** Cet *a priori* est uniforme sur la plage x_1 et x_2 et gaussien à son extrémité supérieure. σ est la largeur caractéristique (écart-type) de l'aile gaussienne. On utilise la notation $UG(x_1, x_2, \sigma)$ pour le désigner. Il est de nature informative.

$$\forall\ x \in\ [x_1, x_2],\ P(x) = C$$
$$\forall\ x \in\ ]-\infty, x_1[= 0$$
$$\forall\ x \in\ ]x_2, +\infty[= C\, exp(-\tfrac{1}{2}(x - x_2)^2/\sigma^2)$$

avec C la constante de normalisation. $C = \left((x_2 - x_1) + \frac{1}{2}\sqrt{2\pi}\sigma\right)^{-1}$.

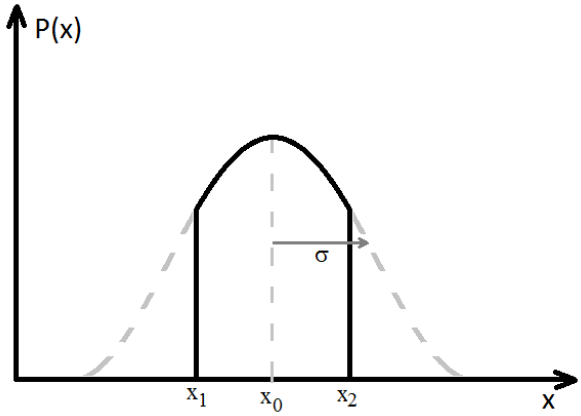


A priori Gaussien-tronqué (GT). Cet a priori est Gaussien sur la plage x_1 and x_2 et nul ailleurs. x_0 et σ correspondent respectivement à la valeur du maximum de densité de probabilité et l'écart-type de la gaussienne. Il est de nature informative.

$$\forall\ x \in [x_1, x_2],\ P(x) \propto exp(-(x - x_0)^2/\sigma^2)$$
$$\forall\ x \in]-\infty, x_1[\cup]x_2, +\infty[,\ P(x) = 0$$

.

Cet *a priori* a été utilisé sous sa forme non normalisée. Ceci n'est aucunement pénalisant à partir du moment où nous ne souhaitons pas faire de comparaison de modèle avec une méthode d'échantillonnage MCMC (le calcul du *global likelihood* nécessite d'utiliser des *a priori* normalisés). On utilise la notation $GT(x_0, \sigma, x_1, x_2)$.



Annexe B

ANNEXE : tables de résultats (paramètres *a posteriori*) pour HD49933

Degré l	Fréquence (μHz)	Intervalle à 1σ	Intervalle à 2σ	Hauteur/Bruit ratio
0	1122.09	+2.37/-4.25	+6.33/-13.6	0.5
0	1206.25	+2.23/-5.57	+3.92/-7.26	0.6
0	1288.97	+1.25/-0.90	+3.20/-1.97	0.7
0	1373.34	+0.93/-1.16	+1.70/-2.50	1.3
0	1460.14	+0.65/-0.80	+1.22/-2.15	1.4
0	1544.69	+0.85/-0.95	+1.76/-2.05	1.2
0	1631.10	+0.59/-0.69	+1.17/-1.59	2.4
0	1714.49	+1.06/-1.17	+2.00/-2.21	1.9
0	1799.75	+0.84/-1.07	+1.62/-2.62	2.4
0	1884.82	+0.88/-1.26	+1.63/-2.65	2.4
0	1972.73	+0.67/-0.71	+1.30/-1.59	2.3
0	2057.82	+3.22/-1.34	+4.72/-2.24	1.9
0	2147.10	+0.76/-0.83	+1.65/-1.99	1.4
0	2236.46	+2.11/-2.83	+3.85/-5.44	0.9
0	2322.10	+2.22/-3.03	+4.08/-5.64	0.7
0	2408.56	+1.97/-2.33	+3.75/-4.81	0.5
0	2495.76	+3.34/-3.21	+7.00/-6.23	0.3
0	2579.85	+4.70/-3.29	+7.61/-10.7	0.3
0	2662.76	+6.01/-4.99	+10.8/-9.23	0.2
1	1079.73	+1.27/-1.27	+3.22/-5.84	0.4
1	1161.54	+0.85/-0.91	+1.67/-2.25	0.6
1	1244.63	+1.02/-1.17	+2.39/-2.66	0.7
1	1328.34	+0.70/-0.65	+1.40/-1.26	1.2
1	1414.93	+0.57/-0.58	+1.22/-1.28	1.4
1	1500.54	+0.70/-0.78	+1.36/-1.62	1.2
1	1586.62	+0.48/-0.49	+0.98/-1.02	2.3
1	1670.48	+0.57/-0.58	+1.16/-1.23	1.9
1	1755.30	+0.51/-0.53	+1.06/-1.05	2.3
1	1840.68	+0.49/-0.50	+0.98/-1.02	2.3
1	1928.13	+0.50/-0.51	+1.00/-1.05	2.3
1	2014.38	+0.54/-0.54	+1.07/-1.09	1.8
1	2101.58	+0.67/-0.72	+1.42/-1.56	1.4
1	2190.81	+0.90/-0.90	+1.88/-1.92	0.8
1	2277.89	+1.16/-1.14	+2.33/-2.37	0.7
1	2362.76	+1.61/-1.61	+3.39/-3.48	0.5
1	2450.35	+2.26/-2.71	+4.58/-7.01	0.3
1	2539.49	+1.50/-4.35	+3.24/-11.2	0.3
1	2625.09	+2.92/-3.46	+5.02/-6.92	0.2

TABLE B.1 – Table de fréquences de HD49933 (IRa-01 + LRa-01) pour le $l = 0$ et $l = 1$.

Degré l	Fréquence (μHz)	Intervalle à 1σ	Intervalle à 2σ	Hauteur/Bruit ratio
2	1115.64	+6.64/-6.64	+11.8/-17.2	0.5
2	1199.91	+4.08/-5.04	+8.77/-11.3	0.6
2	1287.24	+3.51/-3.77	+7.53/-8.65	0.7
2	1369.60	+2.48/-3.11	+5.25/-7.79	1.3
2	1455.42	+2.33/-1.94	+5.49/-4.07	1.4
2	1541.54	+3.07/-4.47	+5.94/-9.48	1.2
2	1626.30	+2.79/-3.00	+5.16/-6.17	2.4
2	1712.67	+2.83/-2.44	+5.23/-4.75	1.9
2	1794.39	+2.73/-2.10	+6.47/-3.91	2.4
2	1881.83	+2.70/-2.02	+5.27/-4.08	2.4
2	1965.19	+2.01/-1.74	+4.78/-3.55	2.3
2	2060.22	+2.88/-5.17	+4.57/-7.40	1.9
2	2140.32	+3.25/-2.67	+7.02/-5.08	1.4
2	2230.68	+5.10/-2.91	+9.99/-5.89	0.9
2	2316.86	+4.35/-3.35	+9.06/-8.24	0.7
2	2403.52	+3.97/-4.56	+7.77/-11.1	0.5
2	2491.50	+4.96/-5.16	+10.3/-11.2	0.3
2	2576.51	+5.03/-7.94	+9.92/-18.6	0.3
2	2656.76	+6.23/-5.62	+13.6/-12.8	0.2

TABLE B.2 – Table de fréquences de HD49933 (IRa-01 + LRa-01) pour le $l = 2$.

Degré l	Fréquence (μHz)	Intervalle à 1σ	Intervalle à 2σ	Hauteur/Bruit ratio
0	1.34	+0.28/-0.31	3.24	+3.54/-2.04
0	1.49	+0.22/-0.22	3.21	+2.19/-1.34
0	2.12	+0.26/-0.24	6.04	+3.60/-2.42
0	2.21	+0.16/-0.16	3.79	+1.12/-0.93
0	2.31	+0.18/-0.17	3.90	+1.39/-1.01
0	2.70	+0.17/-0.17	6.59	+1.97/-1.55
0	3.14	+0.16/-0.15	4.66	+0.99/-0.89
0	3.36	+0.15/-0.15	6.94	+1.23/-1.08
0	3.72	+0.15/-0.15	7.06	+1.11/-1.01
0	3.29	+0.14/-0.14	5.72	+0.90/-0.82
0	3.13	+0.15/-0.14	5.52	+1.10/-1.02
0	2.90	+0.14/-0.14	6.05	+1.21/-1.05
0	2.62	+0.14/-0.14	6.61	+1.69/-1.40
0	2.28	+0.14/-0.15	8.52	+2.14/-1.81
0	2.18	+0.15/-0.15	9.40	+2.48/-2.02
0	1.99	+0.16/-0.16	11.0	+3.36/-2.64
0	1.63	+0.19/-0.19	12.7	+6.71/-4.21
0	1.22	+0.22/-0.22	7.52	+8.72/-4.32
0	1.41	+0.19/-0.20	13.1	+7.38/-5.24
1	1.63	+0.34/-0.38	3.24	+3.54/-2.04
1	1.82	+0.26/-0.28	3.21	+2.19/-1.34
1	2.58	+0.31/-0.30	6.04	+3.60/-2.42
1	2.70	+0.19/-0.20	3.79	+1.12/-0.93
1	2.82	+0.22/-0.21	3.90	+1.39/-1.01
1	3.30	+0.21/-0.21	6.59	+1.97/-1.55
1	3.83	+0.19/-0.19	4.66	+0.99/-0.89
1	4.10	+0.19/-0.19	6.94	+1.23/-1.08
1	4.54	+0.18/-0.18	7.06	+1.11/-1.01
1	4.01	+0.17/-0.17	5.72	+0.90/-0.82
1	3.82	+0.18/-0.17	5.52	+1.10/-1.02
1	3.54	+0.16/-0.17	6.05	+1.21/-1.05
1	3.19	+0.18/-0.18	6.61	+1.69/-1.40
1	2.78	+0.17/-0.18	8.52	+2.14/-1.81
1	2.66	+0.18/-0.18	9.40	+2.48/-2.02
1	2.43	+0.19/-0.20	11.0	+3.36/-2.64
1	1.99	+0.23/-0.23	12.7	+6.71/-4.21
1	1.49	+0.28/-0.27	7.52	+8.72/-4.32
1	1.73	+0.23/-0.25	13.1	+7.38/-5.24

TABLE B.3 – Table d'amplitudes et de largeurs pour HD49933 (IRa-01 + LRa-01) pour le $l = 0$ et $l = 1$.

Degré l	Amplitude (ppm)	Intervalle à 1σ	Largeur (μHz)	Intervalle à 1σ
2	0.98	+0.20/-0.23	3.24	+3.54/-2.04
2	1.09	+0.16/-0.16	3.21	+2.19/-1.34
2	1.54	+0.19/-0.18	6.04	+3.60/-2.42
2	1.61	+0.12/-0.12	3.79	+1.12/-0.93
2	1.68	+0.13/-0.12	3.90	+1.39/-1.01
2	1.97	+0.13/-0.12	6.59	+1.97/-1.55
2	2.28	+0.12/-0.11	4.66	+0.99/-0.89
2	2.44	+0.11/-0.11	6.94	+1.23/-1.08
2	2.71	+0.11/-0.11	7.06	+1.11/-1.01
2	2.39	+0.10/-0.10	5.72	+0.90/-0.82
2	2.28	+0.11/-0.10	5.52	+1.10/-1.02
2	2.11	+0.10/-0.10	6.05	+1.21/-1.05
2	1.91	+0.10/-0.10	6.61	+1.69/-1.40
2	1.66	+0.10/-0.11	8.52	+2.14/-1.81
2	1.59	+0.11/-0.11	9.40	+2.48/-2.02
2	1.45	+0.12/-0.12	11.0	+3.36/-2.64
2	1.19	+0.14/-0.14	12.7	+6.71/-4.21
2	0.89	+0.16/-0.16	7.52	+8.72/-4.32
2	1.03	+0.14/-0.14	13.1	+7.38/-5.24

TABLE B.4 – Table d'amplitudes et de largeurs pour HD49933 (IRa-01 + LRa-01) pour le $l = 2$.

Paramètre	médian	$1\sigma_+$ / $1\sigma_-$	$2\sigma_+$ / $2\sigma_-$
i (°)	27.3	8.7 / 9.6	20.9 / 28.5
ν_s (μ Hz)	3.2	0.9 / 0.8	1.6 / 1.4
$V_{l=1}$	1.49	0.08 / 0.09	0.16 / 0.18
$V_{l=2}$	0.53	0.00 / 0.00	0.00 / 0.00

TABLE B.5 – Table regroupant les paramètres globaux des modes de HD49933 (IRa-01 + LRa-01).

Annexe C

ANNEXE : tables de résultats (paramètres *a posteriori*) pour HD181420

Degré l	Fréquence (μHz)	intervalle à 1σ	intervalle à 2σ	Hauteur/Bruit ratio
0	1053.54	+0.33/-1.28	+0.78/-2.74	1.2
0	1122.95	+1.02/-0.98	+2.27/-2.36	0.7
0	1196.34	+0.36/-0.39	+0.88/-0.84	1.2
0	1272.69	+0.96/-1.28	+2.11/-2.48	0.7
0	1349.24	+0.63/-0.77	+1.37/-1.66	1.1
0	1423.83	+0.57/-0.75	+1.12/-1.92	1.5
0	1496.33	+1.20/-1.12	+2.57/-2.29	1.0
0	1572.58	+1.15/-1.08	+2.29/-2.06	1.5
0	1647.37	+1.54/-1.54	+3.02/-2.99	1.1
0	1721.60	+1.31/-1.07	+2.66/-2.11	1.0
0	1799.81	+1.46/-1.64	+2.87/-3.36	0.9
0	1878.23	+2.53/-2.86	+4.29/-5.73	0.5
0	1952.67	+1.44/-1.51	+3.71/-4.72	0.7
0	2025.95	+2.39/-2.04	+9.16/-4.37	0.4
0	2103.84	+1.64/-1.64	+3.29/-6.68	0.6
0	2181.85	+1.52/-5.67	+6.29/-9.88	0.3
0	2258.26	+5.04/-8.97	+10.2/-15.8	0.2
0	2335.09	+4.57/-9.15	+10.6/-19.1	0.5
0	2409.38	+4.13/-5.14	+6.79/-16.0	0.5
0	2480.00	+4.15/-2.00	+11.5/-4.84	0.5
1	1004.56	+0.48/-0.54	+9.57/-1.57	1.2
1	1081.42	+3.35/-1.02	+6.96/-1.67	0.7
1	1159.20	+1.72/-1.08	+4.22/-1.92	1.2
1	1233.70	+0.99/-1.75	+2.08/-3.54	0.7
1	1307.69	+0.77/-0.91	+1.78/-2.00	1.1
1	1385.18	+0.63/-0.68	+1.32/-1.45	1.5
1	1457.46	+1.27/-1.23	+2.67/-2.38	1.0
1	1535.44	+0.69/-0.75	+1.40/-1.68	1.5
1	1609.41	+0.98/-1.09	+1.89/-2.17	1.1
1	1683.08	+1.11/-1.13	+2.29/-2.29	1.0
1	1760.15	+0.98/-1.23	+2.24/-2.68	0.9
1	1836.06	+2.05/-2.05	+4.26/-4.26	0.5
1	1912.20	+1.08/-1.20	+2.46/-2.91	0.7
1	1987.85	+2.04/-2.42	+4.24/-6.55	0.4
1	2061.33	+1.08/-1.47	+2.36/-3.78	0.6
1	2138.70	+6.06/-2.37	+11.6/-6.06	0.3
1	2220.70	+3.62/-6.03	+9.21/-21.3	0.2
1	2291.72	+6.34/-5.92	+16.2/-7.40	0.5
1	2374.33	+0.95/-5.86	+3.22/-11.7	0.5
1	2448.73	+1.47/-3.10	+6.21/-6.75	0.5

TABLE C.1 – Table de fréquences de HD181420 pour le $l = 0$ et $l = 1$.

Degré l	Fréquence (μHz)	Intervalle à 1σ	Intervalle à 2σ	Hauteur/Bruit ratio
2	1047.85	+3.85/-4.06	+6.22/-6.55	1.2
2	1117.98	+3.18/-5.36	+5.42/-9.43	0.7
2	1194.72	+1.11/-3.84	+2.55/-6.79	1.2
2	1266.51	+3.98/-4.68	+6.52/-8.47	0.7
2	1344.92	+2.63/-4.13	+4.61/-9.44	1.1
2	1416.51	+3.23/-1.95	+5.51/-5.92	1.5
2	1489.89	+3.65/-3.60	+6.00/-6.96	1.0
2	1565.40	+4.33/-4.06	+6.75/-7.02	1.5
2	1639.49	+3.43/-3.33	+6.85/-6.85	1.1
2	1719.10	+1.72/-2.51	+3.21/-5.64	1.0
2	1794.58	+2.74/-2.96	+4.91/-5.75	0.9
2	1871.53	+3.51/-3.45	+6.72/-7.15	0.5
2	1945.47	+4.17/-3.67	+7.07/-8.94	0.7
2	2021.03	+3.57/-3.20	+8.34/-6.92	0.4
2	2096.67	+3.74/-3.58	+6.85/-8.49	0.6
2	2173.89	+4.47/-4.66	+8.46/-10.8	0.3
2	2250.45	+7.25/-8.63	+13.5/-16.4	0.2
2	2328.52	+6.95/-9.77	+13.2/-21.3	0.5
2	2404.42	+4.60/-6.75	+9.20/-19.0	0.5
2	2474.47	+4.92/-5.40	+13.0/-11.1	0.5

Table C.2 – Table de fréquences de HD181420 pour le $l = 2$.

Degré l	Fréquence (μHz)	Intervalle à 1σ	Intervalle à 2σ	Hauteur/Bruit ratio
0	1.87	+0.29/-0.28	1.40	+1.85/-0.74
0	1.93	+0.30/-0.31	2.80	+2.06/-1.36
0	2.02	+0.25/-0.25	1.86	+1.03/-0.81
0	2.17	+0.27/-0.27	4.04	+2.46/-1.57
0	2.62	+0.24/-0.24	3.91	+1.58/-1.24
0	2.80	+0.22/-0.22	3.38	+1.11/-0.83
0	3.51	+0.22/-0.23	7.96	+2.00/-1.67
0	3.50	+0.20/-0.20	5.67	+1.35/-1.10
0	3.48	+0.22/-0.22	7.56	+1.76/-1.46
0	3.33	+0.21/-0.21	7.87	+1.99/-1.64
0	3.10	+0.23/-0.22	7.98	+2.87/-2.27
0	2.78	+0.26/-0.27	12.8	+4.70/-4.15
0	2.45	+0.24/-0.24	6.28	+3.37/-2.63
0	2.25	+0.29/-0.26	9.30	+6.69/-3.78
0	1.89	+0.23/-0.25	4.60	+3.30/-1.93
0	1.23	+0.34/-0.33	3.51	+8.30/-2.63
0	1.44	+0.44/-0.50	11.0	+18.2/-10.1
0	0.68	+0.34/-0.33	0.25	+13.0/-0.22
0	0.92	+0.29/-0.38	1.56	+5.11/-1.13
0	1.14	+0.28/-0.34	2.41	+3.54/-1.93
1	2.17	+0.34/-0.31	1.40	+1.85/-0.74
1	2.23	+0.35/-0.36	2.80	+2.06/-1.36
1	2.34	+0.29/-0.29	1.86	+1.03/-0.81
1	2.52	+0.32/-0.31	4.04	+2.46/-1.57
1	3.04	+0.29/-0.29	3.91	+1.58/-1.24
1	3.25	+0.25/-0.26	3.38	+1.11/-0.83
1	4.07	+0.27/-0.27	7.96	+2.00/-1.67
1	4.06	+0.23/-0.23	5.67	+1.35/-1.10
1	4.03	+0.26/-0.25	7.56	+1.76/-1.46
1	3.86	+0.24/-0.24	7.87	+1.99/-1.64
1	3.60	+0.27/-0.27	7.98	+2.87/-2.27
1	3.22	+0.30/-0.32	12.8	+4.70/-4.15
1	2.85	+0.27/-0.28	6.28	+3.37/-2.63
1	2.61	+0.33/-0.31	9.30	+6.69/-3.78
1	2.19	+0.27/-0.29	4.60	+3.30/-1.93
1	1.43	+0.39/-0.38	3.51	+8.30/-2.63
1	1.67	+0.51/-0.58	11.0	+18.2/-10.1
1	0.78	+0.40/-0.38	0.25	+13.0/-0.22
1	1.07	+0.34/-0.44	1.56	+5.11/-1.13
1	1.32	+0.33/-0.40	2.41	+3.54/-1.93

TABLE C.3 – Table d'amplitudes et de largeurs pour HD181420 pour le $l = 0$ et $l = 1$.

Degré l	Amplitude (ppm)	Intervalle à 1σ	Largeur (μHz)	Intervalle à 1σ
2	1.36	+0.21/-0.20	1.40	+1.85/-0.74
2	1.40	+0.22/-0.22	2.80	+2.06/-1.36
2	1.47	+0.18/-0.18	1.86	+1.03/-0.81
2	1.58	+0.20/-0.20	4.04	+2.46/-1.57
2	1.91	+0.17/-0.18	3.91	+1.58/-1.24
2	2.04	+0.16/-0.16	3.38	+1.11/-0.83
2	2.56	+0.16/-0.16	7.96	+2.00/-1.67
2	2.55	+0.15/-0.14	5.67	+1.35/-1.10
2	2.53	+0.16/-0.16	7.56	+1.76/-1.46
2	2.43	+0.15/-0.15	7.87	+1.99/-1.64
2	2.26	+0.17/-0.16	7.98	+2.87/-2.27
2	2.02	+0.19/-0.20	12.8	+4.70/-4.15
2	1.79	+0.17/-0.18	6.28	+3.37/-2.63
2	1.64	+0.21/-0.19	9.30	+6.69/-3.78
2	1.37	+0.17/-0.18	4.60	+3.30/-1.93
2	0.90	+0.25/-0.24	3.51	+8.30/-2.63
2	1.05	+0.32/-0.36	11.0	+18.2/-10.1
2	0.49	+0.24/-0.24	0.25	+13.0/-0.22
2	0.67	+0.21/-0.27	1.56	+5.11/-1.13
2	0.83	+0.20/-0.25	2.41	+3.54/-1.93

TABLE C.4 – Table d'amplitudes et de largeurs pour HD181420 pour le $l = 2$.

Paramètre	médian	$1\sigma_+$ / $1\sigma_-$	$2\sigma_+$ / $2\sigma_-$
i (°)	33.7	5.5 / 5.9	10.9 / 18.2
ν_s (μ Hz)	4.5	0.7 / 0.9	2.4 / 1.6
$V_{l=1}$	1.34	0.11 / 0.12	0.20 / 0.24
$V_{l=2}$	0.53	0.00 / 0.00	0.00 / 0.00

TABLE C.5 – Table regroupant les paramètres globaux des modes de HD181420.

Résumé : L'héliosismologie a permis de mieux comprendre la structure interne du Soleil et plus généralement de faire des progrès en physique fondamentale. Depuis quelques années, la sensibilité instrumentale a été grandement accrue et il est désormais possible d'étudier les pulsations des consoeurs du Soleil.
Le travail de thèse a essentiellement consisté en l'étude des étoiles de type solaire observées par la mission CoRoT du CNES. Un autre satellite, Kepler (NASA), a été lancé récemment. Plus ambitieux que CoRoT, il recherche depuis plus d'un an des exoplanètes et étudie les pulsations de plusieurs centaines de milliers d'étoiles. Ces données ont aussi été analysées durant ma thèse.
L'analyse des étoiles de type solaire s'est révélée délicate, les pulsations étant d'intensité plus faible qu'attendues et les propriétés intrinsèques de ces pulsations pour les étoiles observées rendant l'analyse difficile. Dans un premier temps le travail de thèse a consisté à trouver un moyen robuste d'extraire un maximum d'information des courbes de lumières de CoRoT. C'est naturellement que nous nous sommes orientés vers une approche Bayésienne d'analyse du signal couplée à un algorithme d'échantillonnage par chaine de Markov. Ce type d'approche repose sur un formalisme mathématique permettant d'incorporer toute l'information *a priori* de l'objet étudié et se révèle robuste même si nous ne sommes capables de formuler que peu d'*a priori*.
Dans ce contexte, nous avons abordé l'analyse du signal stellaire sous deux angles. Tout d'abord, quelles hypothèses de travail devons nous utiliser ou ne pas utiliser lorsque l'on souhaite effectuer une analyse détaillée des propriétés des modes stellaires dans des conditions observationelles difficiles ? Cette question pose le problème de la recherche du modèle optimal. Cette recherche du modèle optimal se fait par recherche du meilleur ajustement du spectre de puissance des étoiles observées. Ce travail montre qu'il est possible d'identifier les modes pulsations des étoiles malgré des rapports signal sur bruit très faibles, comme c'est le cas pour deux étoiles observées par CoRoT : HD49933 et HD181420.
Pour les données Kepler (et dans le futur pour des missions tels que PLATO), nous sommes confrontés à une masse importante de données qu'il nous faut analyser de manière fiable et automatique : comment pouvons nous procéder dans ce cas ? Ce travail propose une réponse : un algorithme d'ajustement automatique est présenté et appliqué au spectre de puissance solaire et sur des données stellaires synthétiques. L'application de l'algorithme développé à ces dernières données montre la possibilité d'une identification automatique des modes fiable dans 78% des cas.
Enfin, les résultats des analyses de HD49933 et HD181420 sont interprétés en terme de propriétés physiques et de structure interne. L'existence d'un coeur convectif au sein de ces deux étoiles est confirmée, ainsi que celle d'une fine enveloppe convective externe.

Mots clés : étoiles, statistiques