

La détection de périodicités cachées

THÈSE N° 5296 (2012)

PRÉSENTÉE LE 20 AVRIL 2012
À LA FACULTÉ DES SCIENCES DE BASE
CHAIRE DE STATISTIQUE APPLIQUÉE
PROGRAMME DOCTORAL EN MATHÉMATIQUES

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Jean-Marc NICOLETTI

acceptée sur proposition du jury:

Prof. T. Mountford, président du jury
Prof. S. Morgenthaler, Dr L. Eyer, directeurs de thèse
Prof. A. C. Davison, rapporteur
Prof. G. Molenberghs, rapporteur
Prof. S. Sardy, rapporteur



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2012

Je tiens en premier lieu à remercier mon directeur de thèse, le Professeur Stephan Morgenthaler pour m'avoir offert l'opportunité d'effectuer un travail de recherche dans ce domaine passionnant liant l'astronomie et les statistiques. Ses conseils, sa gentillesse, sa disponibilité et son soutien autant scientifique que moral m'ont permis d'aller au bout de ma thèse.

Je tiens aussi à remercier mon co-directeur de thèse, le Dr Laurent Eyer, pour son soutien, ses conseils, sa gentillesse et sa disponibilité. Cette collaboration m'a donné l'occasion de découvrir le travail effectué par des astronomes et des développeurs dans une projet d'une telle envergure.

Je remercie aussi le Professeur Anthony C. Davison (EPFL, Suisse), le Professeur G. Molenberghs (Universiteit Hasselt, Belgique) et le Professeur Sylvain Sardy (Université de Genève, Suisse) pour avoir accepté de faire partie de mon Jury. Je remercie également le Professeur Thomas Mountford pour avoir officié en tant que président.

Je remercie aussi tous mes collègues, anciens et actuels, de la section de mathématiques avec qui j'ai eu la chance de passer ces belles années.

Je remercie tous mes amis pour leur soutien moral, les sorties et les soirées passées ensemble, ainsi que ces revigorantes sorties à vélo. Je remercie tout particulièrement Grégoire Bovet pour la lecture de ma thèse.

Je remercie infiniment mes parents Maria et Silvano ainsi que mon frère Diego, pour m'avoir donné la possibilité d'en arriver là, pour leur soutien indéfectible depuis toujours, et surtout pendant la période de mes études et de mon travail doctorat.

Je termine par mon amie, Céline, pour son soutien, ses encouragements, ses conseils et sa patience durant ces années de dur labeur.

Lausanne, février 2012
Jean-Marc Nicoletti

Résumé

Ce travail s'inscrit dans le cadre de la mission Gaia, un satellite de l'Agence Spatiale Européenne (ESA). Un des buts de cette mission est d'effectuer une classification des étoiles variables à l'aide de différents paramètres. La périodicité du changement de la magnitude de l'étoile variable est un paramètre important qui doit être déterminé avec une bonne précision. Une des problématiques est l'irrégularité temporelle des observations. Cela perturbe les méthodes classiques de détection de périodicité.

Dans ce travail, différentes méthodes de détection de périodicité ont été appliquées sur des données réelles et sur des données simulées. Les performances de ces méthodes ont été comparées, pour pouvoir sélectionner la méthode la plus efficace. Dans un deuxième temps, nous nous sommes intéressés à la performance asymptotique des estimateurs fournis par l'analyse harmonique ainsi que par la régression non paramétrique. Nous dérivons la distribution asymptotique de ces deux estimateurs sous l'hypothèse que le modèle des observations contient un signal périodique de période $P_0 > 0$, continu et de carré intégrable sur le domaine $[0, P_0)$, avec des erreurs Gaussiennes corrélées. L'échantillonnage irrégulier vérifie la propriété asymptotique que l'intervalle $[0, P_0)$ peut être entièrement observé. Nous avons dérivé une approximation de la distribution asymptotique de l'estimateur de la régression non paramétrique dans le cas où le signal périodique de période $P_0 > 0$ est constant par morceau sur l'intervalle $[0, P_0)$.

Nous nous sommes aussi intéressés au cas où l'échantillonnage ne vérifie pas la propriété asymptotique que le signal peut être complètement observé. En considérant un cas particulier de ce type, nous avons dérivé la distribution asymptotique de l'estimateur de la régression non paramétrique, sous l'hypothèse que le modèle des observations contient un signal périodique intégrable et continu, avec des erreurs indépendantes et identiquement distribuées.

Mots-clés : échantillonnage irrégulier ; signal périodique ; analyse harmonique ; régression non paramétrique ; distribution asymptotique.

Abstract

This thesis is a small part of the preparation of the launch of the Gaia mission, a satellite of the European Space Agency (ESA). One of the goals of the mission is to perform a classification among variable stars considering different attributes. Periodic behavior in the observed light curve is such an attribute. It is of importance to determine these hidden cycles with as high an accuracy as possible. A key difficulty is connected to the fact that observations are taken at irregularly distributed time points. Classical methods of frequency analysis do not work in this situation.

In a first step, we made a catalogue of potential solutions and applied them to real and simulated data. The performances have been compared in order to select the best method. In a second step, we considered the asymptotic performance of estimators based on regression methods. We derived the asymptotic distribution under the hypothesis that the observation model contains a periodic signal with period $P_0 > 0$, continuous and square integrable on $[0, P_0)$, with Gaussian correlated errors. The irregular sampling has to satisfy the asymptotic property that the whole interval $[0, P_0)$ can be observed. We have also considered the special case of a periodic signal with period $P_0 > 0$, piecewise constant on $[0, P_0)$ and the asymptotic distribution of the estimator.

If an irregular sampling scheme does not satisfy the asymptotic property that the signal can be fully observed, the period can still be estimated reliably. We determined the asymptotic distribution of an estimator in a particular situation and under the assumption that the observation model contains a periodic signal with period $P_0 > 0$, continuous and square integrable on $[0, P_0)$, and is observed with additive errors that are independent and identically distributed.

Keywords : Irregular sampling ; periodicity ; harmonic analysis ; nonparametric regression ; asymptotic distribution.

Table des matières

Résumé	3
Abstract	7
1 Introduction	11
1.1 Motivation	11
1.2 La mission Gaia	12
1.3 Les étoiles variables	13
1.4 Les observations et la courbe lumière	15
1.5 Séries temporelles	15
1.6 Processus stationnaires	16
1.7 Périodogramme	17
1.8 L'autocorrélation	20
1.9 Théorie de Fourier	22
1.9.1 Théorie des séries de Fourier	23
1.9.2 Sommes partielles d'une série de Fourier	24
1.9.3 La transformée de Fourier	26
1.10 Echantillonnage	26
1.10.1 Echantillonnage régulier	27
1.10.2 Echantillonnage irrégulier	27
1.11 La recherche de fréquence dans la littérature scientifique	29
2 Méthodes de recherche de fréquences cachées	33
2.1 L'analyse harmonique	33
2.2 Méthode de Deeming	36
2.3 Remarques sur les méthodes de l'analyse harmonique et de Deeming	37
2.4 Estimation par régression non paramétrique	38
2.5 Mariage entre l'analyse harmonique et la régression non paramétrique	39
2.6 Récapitulatif des méthodes de régressions	41
3 Application des méthodes sur la partie A du catalogue des étoiles variables de la mission Hipparcos	43
3.1 Modèle des observations	43
3.1.1 Modèle avec des erreurs corrélées	44
3.2 Comparaison des méthodes : analyse harmonique classique et pondérée, autocorrélation, régression non paramétrique et périodogramme en amplitude (Deeming)	44
3.2.1 Comparaison entre l'analyse harmonique et l'analyse harmonique pondérée	45
3.2.2 Comparaison des performances des méthodes de l'analyse harmonique, l'autocorrélation, la régression non paramétrique et la méthode de Deeming	45
3.2.3 Analyse de la performance des méthodes sur les données de la partie A du catalogue des étoiles variables de la mission Hipparcos	45
3.2.4 Conclusion	56

4	Distributions asymptotiques	61
4.1	Historique de la théorie asymptotique	61
4.1.1	Signal sinusoidal, échantillonnage régulier et bruit Gaussien indépendant	62
4.1.2	Signal sinusoidal, échantillonnage régulier et bruit Gaussien corrélé	63
4.2	Echantillonnage irrégulier vérifiant la condition de couverture de la période	64
4.2.1	Signal périodique quelconque, échantillonnage irrégulier et bruit Gaussien indépendant	64
4.2.2	Propriétés asymptotiques de l'estimateur $\hat{\omega}_n^{\text{AH}}$	64
4.2.3	Distribution asymptotique de l'estimateur $\hat{\omega}_n^{\text{NP}}$	87
4.3	Echantillonnage irrégulier ne vérifiant pas la condition de couverture de la période	93
4.3.1	Distribution asymptotique de l'estimateur $\hat{\omega}_n^{\text{NP}}$	94
4.3.2	Cas particulier de ce modèle d'échantillonnage	95
4.4	Exemples	96
4.4.1	Variance asymptotique de $\hat{\omega}_n^{\text{NP}}$ pour un signal boîte	96
4.4.2	Varier la densité spectrale	97
5	Simulations	101
5.1	Modèles des données simulées	101
5.1.1	Signaux périodiques	101
5.1.2	Modèles d'échantillonnage	102
5.2	Comparaison de la performances des méthodes : analyse harmonique, régression non paramétrique et périodogramme en amplitude (Deeming)	104
5.2.1	Modèles d'échantillonnage avec la propriété de couverture de la période	106
5.2.2	Modèles d'échantillonnage ne vérifiant pas la couverture de la période	116
5.3	Simulation de données avec erreurs gaussiennes corrélées	124
5.4	Simulation de données avec le signal g_5 , avec erreurs gaussiennes indépendantes	126
5.5	Comparaison de la variance asymptotique avec la variance estimée	128
5.6	Conclusion	129
6	Conclusion	133
7	Appendix	135
	Références	138

Chapitre 1

Introduction

1.1 Motivation

Dans le monde qui nous entoure, la plupart des phénomènes physiques, météorologiques, astronomiques ou encore sismologiques, ont un comportement structuré. Modéliser ce comportement et déterminer les paramètres de ce modèle pour ensuite prédire le comportement de ce phénomène dans le futur fait partie de la science des statistiques. En météorologie, on peut observer par exemple une variation saisonnière des températures. Au contraire en sismologie, un séisme ne peut être prédit à l'avance. Cependant les répliques ont un comportement périodique, qui a justement une structure qui peut être modélisée. Cette modélisation du comportement ne se limite pas aux phénomènes physiques. Par exemple, on peut retrouver des structures dans le comportement alimentaire d'une population en fonction de la période de l'année.

Dans ce travail, le sujet d'application est l'astronomie. La structure qui nous intéresse est la variation de l'intensité lumineuse des étoiles au cours du temps. Pour une certaine population d'étoiles, la variation est récurrente. Pour cette catégorie, l'intensité lumineuse est caractérisée par un paramètre de temps, la période de récurrence de l'intensité lumineuse, et par la forme de la fonction représentant cette intensité. Ce travail s'inscrit dans le cadre du projet Gaia, qui est décrit dans la section 1.2, avec une introduction aux étoiles présentant une variation dans leur intensité lumineuse. Dans le reste de ce chapitre, nous verrons les approches classiques pour le traitement de ces observations, en commençant par la théorie de Fourier, section 1.9, le fondement de l'analyse de la périodicité de fonctions périodiques. La section 1.5 sera consacrée à la théorie des séries temporelles. Cela permettra de poser le cadre du modèle et de présenter des méthodes classiques de la théorie des séries temporelles d'analyses de données liées à la variable de temps. La section 1.10, sur les modèles d'échantillonnage, permettra de présenter les différents modes d'échantillonnages. Avec ces notions théoriques, on pourra passer à la partie 1.11, dédiée à l'historique et à l'évolution des méthodes sur ce sujet.

Le but de cette recherche est de tester différentes méthodes de recherche de fréquences cachées sur des données astronomiques, ainsi que simulées, capables de caractériser les fréquences de manière automatiques, sans intervention humaine. Le chapitre 2 présente les méthodes considérées dans ce travail. Le chapitre 3 compare les performances des différentes méthodes considérées. Ce chapitre permettra de décider quelles méthodes sont les plus performantes sur des données provenant de la mission Hipparcos (1989 - 1993), publiées par l'Agence Spatiale Européenne (ESA) en 1997 ¹. Laurent Eyer, astrophysicien et maître d'enseignement et de recherche à l'observatoire de Genève, nous a fourni toutes ces données d'étoiles dont la période, ainsi que la forme de la fonction représentant la variation de l'intensité lumineuse, sont supposées connues. Le chapitre 4 traite de la distribution asymptotique des estimateurs définis au chapitre 2. Le chapitre 5 compare les deux méthodes, dont les estimateurs ont été étudiés dans le chapitre 4, sur des échantillons de données simulées et de tailles finies.

¹Source : <http://www.rssd.esa.int/index.php?project=HIPPARCOS>

1.2 La mission Gaia

Gaia est un satellite de l'Agence Spatiale Européenne (ESA) qui a pour but principal d'observer et d'enregistrer les données d'environ un milliard d'objets. Par objets, on considère les étoiles, les astéroïdes, les étoiles doubles (deux étoiles gravitant autour du centre de masse du système), les quasars (quasi stellar radio source) et les galaxies. Les mesures de la direction et de la distance de chaque objet permettra de faire une cartographie en trois dimensions d'une part de notre galaxie, la voie lactée. De plus le satellite va mesurer le mouvement de chaque objet, et ce qui donnera une idée des mouvements des composantes de la galaxie (disque, halo, bulbe). Gaia mesurera les propriétés physiques de chaque étoile, comme la température et la composition chimique en surface. Il va aussi mesurer des changements associés à l'intensité lumineuse ou à la couleur des étoiles. Les étoiles ayant de telles changements (luminosité et/ou couleur) sont des étoiles dites variables. Plus une étoile a une couleur rouge, moins sa température est élevée. La vitesse radiale est aussi une des caractéristiques mesurées par le satellite. Il s'agit de la vitesse à laquelle l'objet s'éloigne ou se rapproche de l'observateur (ici le satellite). Cela donne la troisième composante pour la vision dans l'espace. Le but scientifique principal de cette mission est de comprendre et de décrire l'histoire de la formation de la galaxie et son évolution. Les objets qui seront repérés par Gaia sont tous ceux qui ont une magnitude inférieure à $V \sim 20$ mag. La magnitude (mag) est la mesure de l'éclat dans la zone de longueur d'onde correspondante (ou bande photométrique). Historiquement, plus l'objet est brillant plus il a une faible magnitude (qui peut être négative). Ainsi une intensité de 20 mag est extrêmement faible. De plus, il faut distinguer deux types de mesure de la magnitude

1. La magnitude absolue correspond à la luminosité intrinsèque de l'étoile.
2. La magnitude apparente est l'intensité lumineuse observée depuis la Terre, corrigée de l'effet de l'atmosphère.

Le satellite mesurera ainsi la magnitude apparente. Pour pouvoir en déduire la luminosité à partir de la magnitude apparente, il faut connaître la distance entre l'observateur et l'étoile. La règle suivante donne le lien entre la magnitude apparente et la magnitude absolue.

La magnitude apparente d'un objet équivaut à sa magnitude absolue s'il se trouve à une distance de 10 parsecs (32.6 années lumières) de l'observateur.

De plus, une différence de 1 mag correspond à un brillance d'environ 2.512 fois plus élevée. Ainsi un objet de magnitude 2 mag est 2.512 fois plus brillant qu'un objet de magnitude 3 mag. Pour effectuer ces mesures, le satellite sera doté d'un plan focal (Fig. 1.1). Le satellite aura à son bord deux télescopes qui observeront le ciel dans deux directions simultanément, séparées d'un angle de 106.5° . La lumière sera redirigée, par l'intermédiaire de miroirs, sur le plan focal. Il y aura ainsi une superposition des deux prises de vues sur le plan. Pour comprendre l'intérêt de ce système, décrivons le mode de fonctionnement du satellite. Gaia sera placé en orbite autour du soleil au point de Lagrange L2. Les points de Lagrange sont des points où un objet, de masse négligeable vis-à-vis du système considéré, composé de deux corps, paraîtra immobile par rapport à eux. En d'autres termes, cet objet aura une vitesse angulaire égale à la vitesse de rotation des deux corps autour de leur centre de masse. Le point de Lagrange L2 se trouve à 1.5 millions de kilomètres de la Terre et à l'opposé du Soleil. Une fois positionné autour du point L2, le satellite aura son axe de rotation faisant un angle de 45° par rapport à la direction du Soleil. L'axe de rotation effectuera un mouvement de précession complété en 63 jours et un mouvement de rotation du satellite effectué en 6 heures. Le but de ces deux angles de vues est d'observer un objet à un moment donné, puis de normalement pouvoir l'observer à nouveau grâce à la deuxième prise de vue et cela 1 heure 46 minutes après le premier passage, bien que la direction du deuxième point d'observation se sera légèrement décalé par rapport à la direction du premier point d'observation, ceci dû au mouvement de précession (figure 1.2). Les observations prises par les deux télescopes seront superposées sur le plan focal simultanément. Ce plan focal (figure 1.1) est constitué de 106 CCD (de l'anglais : Charged-Couple-Devices, en français dispositif à transfert de charges). Un CCD est un circuit intégré composé de pixels. Le CCD produit un signal dépendant de

la quantité de lumière reçue. Lorsque la lumière d'un objet traverse un CCD, un électron est émis de chaque pixel ayant reçu de la lumière et est transféré de pixel en pixel jusqu'à la fin du CCD, et cela avec une vitesse égale à celle de la vitesse de rotation du satellite. Les électrons arrivés à la fin des CCD sont convertis en signal numérique. Le signal est finalement sommé sur l'ensemble des CCD et fournit une valeur moyenne qui est la magnitude de l'objet. Ces CCD ont principalement les fonctions suivantes :

- Les colonnes 2 et 3 (14 CCD), détectent automatiquement les objets et communiquent les détails du passage de l'objet au CCD suivant.
- Les colonnes 4 à 12 (63 CCD), concernent les mesures astrométriques (positions, distances et mouvements des objets).
- Les colonnes 13 et 14 (7 CCD chacun), mesurent la couleur bleue ou rouge, respectivement, de l'objet, à l'aide d'un spectre à basse résolution (températures, extinctions).
- Les colonnes 15 à 17 (12 CCD), enregistrent le spectre de chaque objet ayant une magnitude apparente supérieure à 17 mag (vitesses radiales).

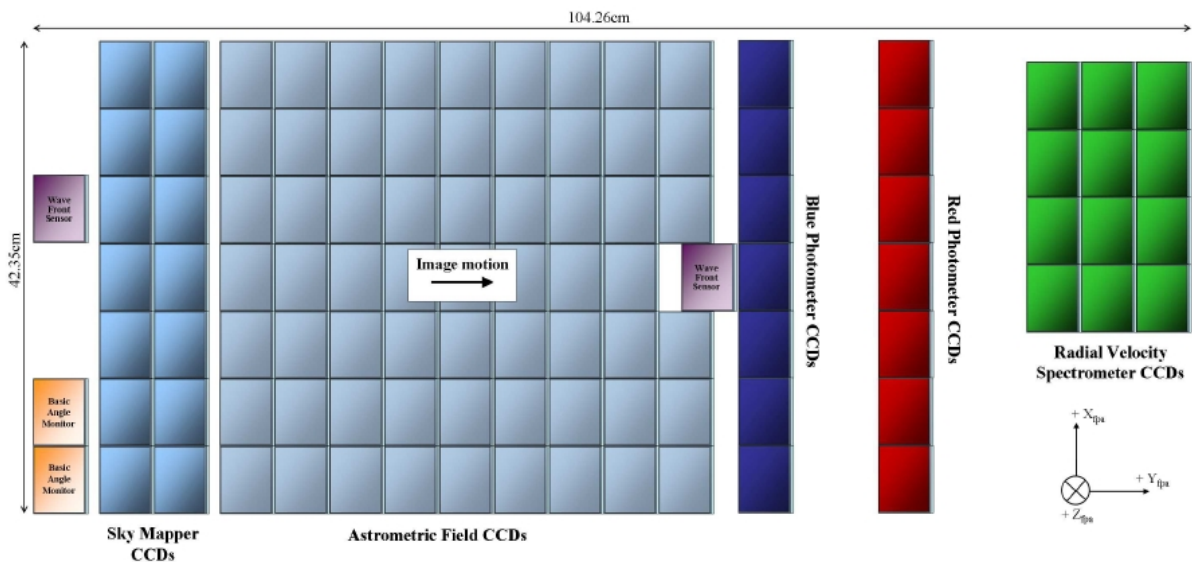


FIG. 1.1 – Plan focal (source : <http://www.rssd.esa.int/index.php?project=Gaia>).

1.3 Les étoiles variables

Une étoile est dite variable lorsque les variations de sa magnitude ou de sa couleur sont détectables sur une échelle de temps de l'ordre d'une vie humaine. Les variations peuvent être périodiques, semi périodiques ou irrégulières avec une échelle de temps variant de quelques minutes à plus de 100 années. Posons $m(t)$ la magnitude d'une étoile, une fonction du temps. L'étoile est dite périodique de période $P > 0$ si sa magnitude $m(t)$ vérifie la relation (1.1)

$$m(t) = m(t + P) \text{ pour tout } t \in \mathbb{R}. \quad (1.1)$$

Ces objets fournissent beaucoup d'informations sur la nature chimique de l'étoile, sa structure interne et surtout permet de comprendre l'évolution des étoiles aux différents stades de leur vie. Les étoiles variables sont recensées dans le catalogue GCVS (General Catalogue of Variables Stars), où elles sont réparties selon les groupes suivants :

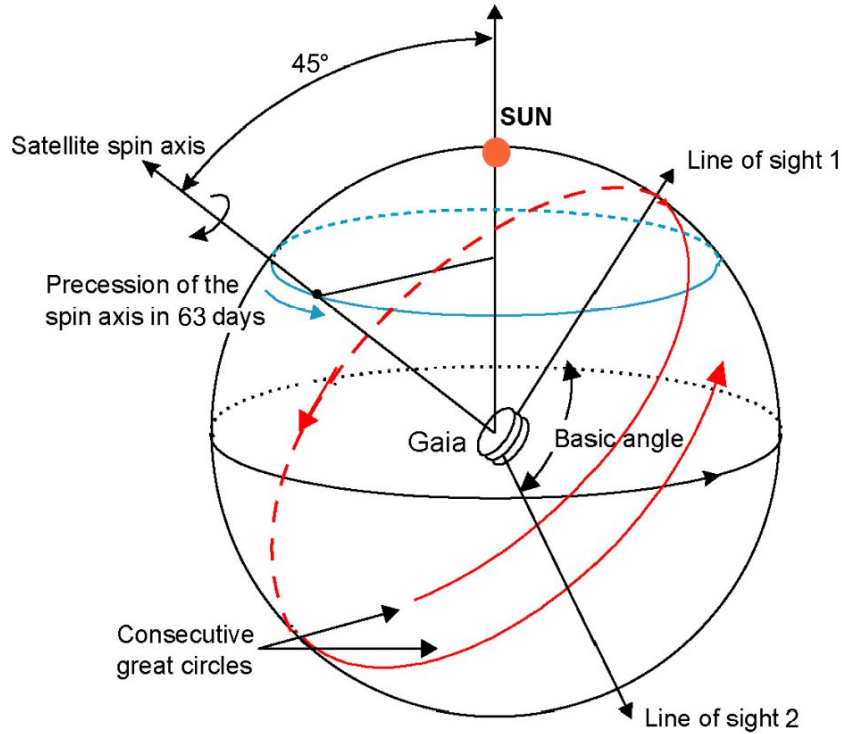


FIG. 1.2 – Schéma des mouvements de rotations et précessions du satellite Gaia (source : <http://www.rssd.esa.int/index.php?project=Gaias>).

- Les variables pulsantes. La surface de l'étoile change périodiquement, dû à des changements physiques internes à l'étoile. La pulsation peut être radiale ou non radiale (ondes transversales de faibles magnitudes se propageant à travers la surface de l'étoile). Par exemple, les Céphéides, les RR Lyrae (RRAB, RRC), les Miras (M), les Dela Scuti (DSCT) ou encore les semi-régulières (SR) sont des étoiles pulsantes avec des périodes très différentes. Les Miras ont des périodes longues (de 80 à 1000 jours), les Céphéides ont des périodes plutôt courtes (de 1 à 70 jours) et les RR Lyrae de très courtes périodes (de ~ 1 h à 1 jour).
- Les variables éruptives. La variabilité est fortement marquée par des changements irréguliers et courts de la magnitude ou par des éruptions de grandes magnitudes. Ces variables sont divisées en plusieurs groupes selon la localisation de l'éruption dans l'étoile (dans la partie haute de l'atmosphère de l'étoile (couronne) ou dans l'atmosphère proche de l'étoile (chromosphère) ou encore à l'intérieur de l'étoile, ainsi que d'autres facteurs additionnels). Par exemple, les variables irrégulières (I) (peu ou pas d'indice de périodicité dans le changement de luminosité), les T Tauri, des étoiles très jeunes, ou encore les S Dauradus, des supergéantes bleues très lumineuses.
- Les variables cataclysmiques. Les étoiles montrant des éruptions causées par des explosions thermonucléaires dans les couches de surface de l'étoiles (Novae) ou profondément, dans leur intérieur (Supernovae).
- Les variables rotatives avec une luminosité à la surface non uniforme ou avec une forme ellipsoïdale. Leur variabilité étant due à la rotation axiale par rapport à un observateur.
- Les binaires à éclipses. Il s'agit d'un système de deux étoiles, en général non variables, qui orbitent autour de leur centre de masse. L'observateur doit être dans le plan orbital. La magnitude est minimale lorsque les deux étoiles sont côte à côte dans la direction de l'observateur. Le premier minimum (de magnitude la plus élevée) apparaît lorsque la plus brillante des deux se trouve derrière l'autre. Il existe plusieurs types de binaires à éclipses, les types E, EA, EB, EW, ELL. La courbe lumière étant différente pour ces différents types. Cette différence provient soit de

la différence de luminosité entre les deux étoiles, soit de la différence de taille, ainsi que de la distance entre elles, ce qui crée des orbites de différentes longueurs.

Des exemples de courbes lumières des types RRAB, RRC, M, DSCT, SR, E, EA, EB, EW, ELL sont présentés dans le chapitre 3 sur l'analyse des méthodes sur les données Hipparcos. Dans la suite de ce travail, nous utiliserons plutôt le nom "objet" que "étoile", car une binaire à éclipse est un groupe de deux étoiles.

1.4 Les observations et la courbe lumière

Les temps d'observation sont en jours juliens tronqués (TJD). Le premier jour julien tronqué a débuté le 24 mai 1968 à minuit. Ces observations sont représentées par graphique (Fig. 1.3) appelé courbe lumière, où la graduation de l'axe des ordonnées (la magnitude) décroît du bas vers le haut. En effet, comme nous l'avons expliqué dans la partie 1.2, une magnitude inférieure correspond à un plus grand éclat.

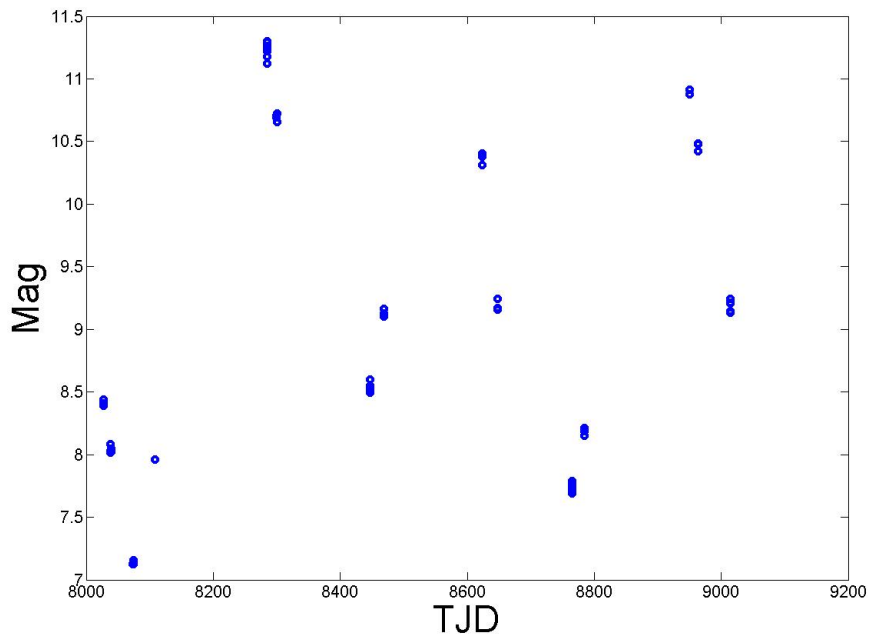


FIG. 1.3 – Observation de la courbe lumière de l'étoile HIP 8 de type Mira (source : Catalogue Hipparcos).

1.5 Séries temporelles

Les séries temporelles sont l'étude d'une certaine quantité dépendante du temps. On les représente comme un couple temps/observation (t_i, y_i) où y_i est la i -ème observation au temps t_i . Un exemple est la figure 1.3 de la courbe lumière observée de l'étoile HIP 8. Soit le modèle en temps continu

$$Y(t) = g(t) + \varepsilon(t), \quad t \in \mathbb{R}, \quad (1.2)$$

où $g(t)$ est le signal sous-jacent et $\varepsilon(t)$ est un processus stochastique Gaussien en temps continu. Les observations aux temps t_1, \dots, t_n conduisent au modèle en temps discret

$$Y_i = g_i + \varepsilon_i, i = 1, \dots, n, \quad (1.3)$$

où $g_i = g(t_i)$ est le signal au temps t_i et $\varepsilon_i = \varepsilon(t_i)$ est l'erreur au temps t_i . L'observation y_i est une réalisation de Y_i .

1.6 Processus stationnaires

Un processus $Y(t)$ continu est dit faiblement stationnaire si :

1. $\mathbb{E}[Y(t)] = \mu$ existe et est constante pour tout $t \in \mathbb{R}$;
2. l'espérance $\mathbb{E}[Y(t)^2]$ est finie ;
3. l'autocovariance $\gamma(s) = \text{Cov}[Y(t), Y(t+s)] = \mathbb{E}[Y(t)Y(t+s)] - \mu^2$ ne dépend que de s .

Soit les variables aléatoires $Y(t_1), \dots, Y(t_n)$, avec $t_i \in \mathbb{R}, i = 1, \dots, n$. Si pour tout $(t_1, \dots, t_n) \in \mathbb{R}^n$ et pour tout décalage $s \in \mathbb{R}$, la loi conjointe de $(Y(t_1), \dots, Y(t_n))$ est identique à la loi conjointe de $(Y(t_1 + s), \dots, Y(t_n + s))$, alors le processus stochastique $Y(t)$ est dit strictement stationnaire. Un processus $Y(t)$ est ergodique si la moyenne temporelle de ses propriétés est égale à la moyenne de tous ses événements. Plus précisément, un processus stationnaire $Y(t)$ est dit ergodique de premier ordre si les moyennes temporelles $\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T Y(t)^k dt$ sont égales aux moments $\mathbb{E}[Y(t)^k]$, $k = 1, 2, \dots$ et $t \in \mathbb{R}$. De plus, un processus stationnaire $Y(t)$ est dit ergodique de deuxième ordre si $\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T (Y(t+s) - \mu)(Y(t) - \mu) dt = \gamma(s)$. Si le processus $Y(t)$ est stationnaire et ergodique de premier et de deuxième ordre, le théorème de Herglotz peut être appliqué (voir par exemple [3]) et garantit l'existence d'une distribution spectrale $F_{sp}(\omega), \omega \in \mathbb{R}$, telle que

$$\gamma(s) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(is\omega) dF_{sp}(\omega). \quad (1.4)$$

Si $F_{sp}(\omega)$ est différentiable pour tout $\omega \in \mathbb{R}$, alors on peut définir la densité spectrale $f_{sp}(\omega)$ comme

$$f_{sp}(\omega) = \int_{-\infty}^{\infty} \gamma(s) \exp(-is\omega) ds. \quad (1.5)$$

L'autocovariance est donc la transformée inverse de la densité spectrale $f_{sp}(\omega)$, si elle existe,

$$\gamma(s) = \int_{-\infty}^{+\infty} f_{sp}(\omega) \exp(is\omega) d\omega. \quad (1.6)$$

Un modèle simple d'un processus stochastique $\varepsilon(t)$ de (1.2) est le bruit blanc Gaussien, qui est tel que $\varepsilon(t)$ et $\varepsilon(u)$ sont indépendants pour tout $t \neq u$. A présent, considérons un ensemble de temps t_1, \dots, t_n , régulièrement espacés, $t_{i+1} - t_i = 1, i = 1, \dots, n-1$, auxquels on a observé le processus stochastique $Y(t)$. L'autocovariance $\gamma(k)$ ne peut être estimée qu'aux décalages $|k| \in \{1, \dots, n-1\}$, par

$$\hat{\gamma}(k) = \frac{1}{n} \sum_{0 \leq l < n, 0 \leq k+l < n} (y_l - \bar{y})(y_{l+k} - \bar{y}). \quad (1.7)$$

Une estimation de la densité spectrale est donnée par

$$\hat{f}_{sp}(\omega) = \frac{1}{2\pi} \sum_{k=-n+1}^{n-1} \hat{\gamma}(k) \exp(-ik\omega). \quad (1.8)$$

En substituant (1.7) dans (1.8) et en prenant que $\bar{y} = 0$, on obtient

$$\begin{aligned}
\widehat{f}_{sp}(\omega) &= \frac{1}{2\pi} \sum_{k=-n+1}^{n-1} \exp(-ik\omega) \frac{1}{n} \sum_{0 \leq l < n, 0 \leq k+l < n} y_l y_{l+k} \\
&= \frac{1}{2\pi} \sum_{k=-n+1}^{n-1} \exp(-ik\omega) \frac{1}{n} \sum_{0 \leq l < n, 0 \leq k+l < n} y_l y_{l+k} \exp(-il\omega) \exp(il\omega) \\
&= \frac{1}{2\pi} \sum_{k=-n+1}^{n-1} \frac{1}{n} \sum_{0 \leq l < n, 0 \leq k+l < n} y_l y_{l+k} \exp(-i(l+k)\omega) \exp(il\omega) \\
&= \frac{n}{2\pi} \sum_{k=-n+1}^{n-1} \left| \frac{1}{n} \sum_{l=0}^{n-1} y_l \exp(il\omega) \right|^2 \\
&= \frac{1}{2\pi} \frac{n}{n^2} \left| \sum_{l=0}^{n-1} y_l \exp(il\omega) \right|^2, \tag{1.9}
\end{aligned}$$

car si $k = -(n-1)$, alors $k+l = 0$ et $l = n-1$. Si $k = n-1$, alors $k+l = n-1$ et $l = 0$. Ou encore, si $k = -1$, alors $1 \leq l \leq n-1$ et $0 \leq k+l \leq n-2$. Ainsi que, si $k = 1$, alors $0 \leq l \leq n-2$ et $1 \leq k+l \leq n-1$. On voit que l'on retrouve un produit de deux sommes conjuguées, ce qui donne le module au carré. La formule de la ligne (1.9) est appelée le périodogramme, une estimation de la densité spectrale. Dans la littérature statistique, on la nomme plutôt $I(\omega)$ au lieu de $\widehat{f}_{sp}(\omega)$. La figure 2.4 montre le périodogramme appliqué à des observations y_1, \dots, y_n bruitées et régulièrement espacées, ayant comme signal périodique sous-jacent la fonction

$$g(t) = \cos(\omega_0 t) + \sin(\omega_0 t), \tag{1.10}$$

de fréquence $\omega_0 = \frac{\pi}{5}$ (ou de manière équivalente, de période $P_0 = 10$). Ces observations sont représentées dans la figure 1.4.

1.7 Périodogramme

Le périodogramme (1.9), dans le cas de données régulièrement espacées dans le temps, est utilisé pour estimer la densité spectrale mais aussi pour l'estimation de la, ou les, fréquence(s) présente(nt) dans un signal $g(t)$ quelconque. La figure 2.4 représente le périodogramme des données pour le signal périodique (1.10). Les fréquences considérées sont les fréquences de Fourier $\omega_k = \frac{2\pi k}{n}$, $k = 1, \dots, n$. La décision qu'une fréquence est présente dans un signal $g(t)$ donné se base sur un test statistique $T(\omega) = 2 \frac{n}{\sigma^2} \left(\left(\frac{\sum_{j=1}^n Y_j \cos(\omega t_j)}{n} \right)^2 + \left(\frac{\sum_{j=1}^n Y_j \sin(\omega t_j)}{n} \right)^2 \right)$, avec σ^2 supposée connue. Sous l'hypothèse nulle H_0 : les données ne contiennent pas la fréquence ω , $T(\omega) \sim \chi_2^2$. Ce test rejette l'hypothèse nulle, si la p -valeur de la statistique $T(\omega)$ est plus petite que α (= 5%, en général). En effet, dans le cas où les données ne contiennent que du bruit $Y_j = \varepsilon_j$, $j = 1, \dots, n$, iid, alors $\sum_{j=1}^n Y_j \cos(\omega t_j)$ et $\sum_{j=1}^n Y_j \sin(\omega t_j)$ suivent des lois normales $\mathcal{N}(0, \frac{\sigma^2}{n})$. Ainsi $\left(\frac{1}{n} \sum_{j=1}^n Y_j \cos(\omega t_j) \right)^2$ et $\left(\frac{1}{n} \sum_{j=1}^n Y_j \sin(\omega t_j) \right)^2$ sont distribués selon la loi χ_1^2 avec une échelle $\frac{\sigma^2}{n}$. On peut calculer l'espérance et la variance du périodogramme sous l'hypothèse H_0 . Sous l'alternative, en utilisant le fait que $Y_j = g_j + \varepsilon_j$, et en remarquant que

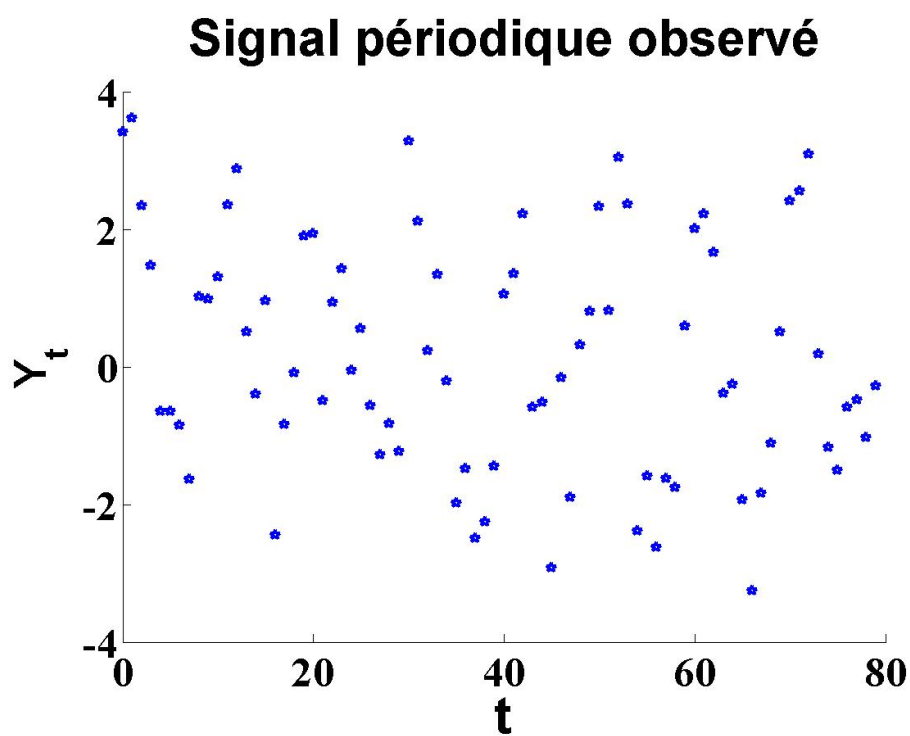


FIG. 1.4 – Données du signal périodique de période $\omega_0 = \frac{\pi}{5}$.

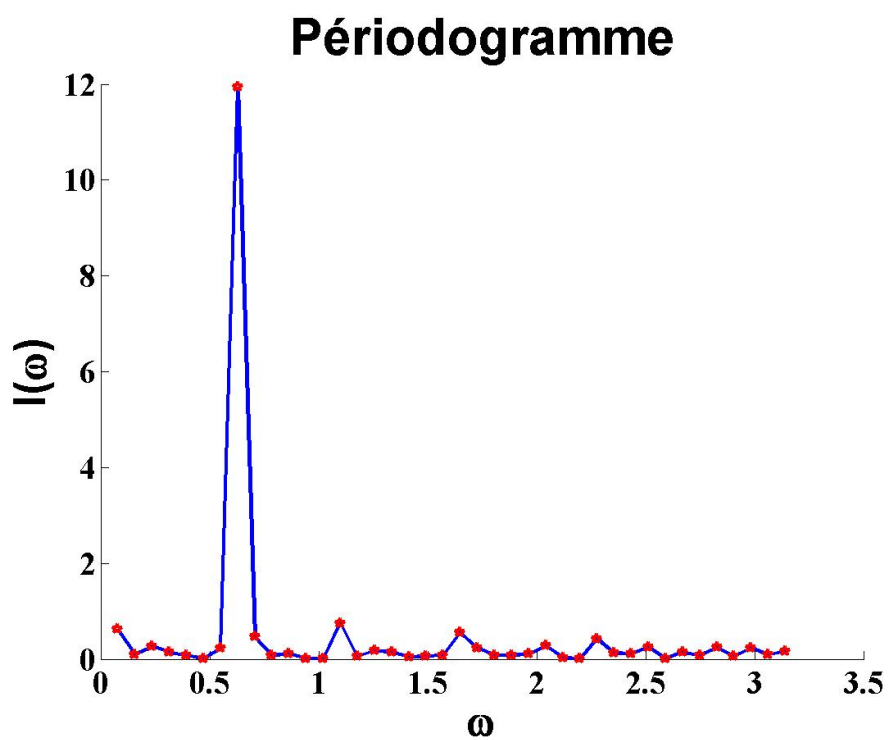


FIG. 1.5 – Périodogramme (1.9) des données pour le signal périodique (1.10).

$$\begin{aligned} \frac{1}{n^2} \left(\sum_{j=1}^n g_j \cos(\omega j) \right)^2 + \frac{1}{n^2} \left(\sum_{j=1}^n g_j \sin(\omega j) \right)^2 &= \frac{1}{n^2} \sum_{j=1}^n \sum_{l=1}^n g_j g_l (\cos(\omega j) \cos(\omega l) + \sin(\omega j) \sin(\omega l)) \\ &= \frac{1}{n^2} \sum_{j=1}^n \sum_{l=1}^n g_j g_l \cos(\omega(j-l)), \end{aligned}$$

on obtient

$$\begin{aligned} \frac{2\pi}{n} \mathbb{E} [I(\omega)] &= \frac{1}{n^2} \left(\sum_{j=1}^n g_j \cos(\omega j) \right)^2 + \frac{1}{n^2} \left(\sum_{j=1}^n g_j \sin(\omega j) \right)^2 + \frac{1}{n^2} \sum_{j=1}^n \sum_{l=1}^n \mathbb{E} [\varepsilon_j \varepsilon_l] \cos(\omega(j-l)) \\ &= \frac{1}{n^2} \left(\sum_{j=1}^n g_j \cos(\omega j) \right)^2 + \frac{1}{n^2} \left(\sum_{j=1}^n g_j \sin(\omega j) \right)^2 + \frac{\sigma^2}{n}. \end{aligned}$$

La variance vaut

$$\begin{aligned} \frac{(2\pi)^2}{n^2} \text{Var} [I(\omega)] &= \frac{1}{n^4} \text{Var} \left[\left(\sum_{j=1}^n (g_j + \varepsilon_j) \cos(\omega j) \right)^2 + \left(\sum_{j=1}^n (g_j + \varepsilon_j) \sin(\omega j) \right)^2 \right] \\ &= \frac{1}{n^4} \mathbb{E} \left[\left(\sum_{j=1}^n (g_j + \varepsilon_j) \cos(\omega j) \right)^4 \right] + \mathbb{E} \left[\left(\sum_{j=1}^n (g_j + \varepsilon_j) \cos(\omega j) \right)^2 \left(\sum_{j=1}^n (g_j + \varepsilon_j) \sin(\omega j) \right)^2 \right] \\ &\quad + \mathbb{E} \left[\left(\sum_{j=1}^n (g_j + \varepsilon_j) \sin(\omega j) \right)^4 \right] - \frac{\sigma^4}{n^2}. \end{aligned}$$

Sous l'hypothèse nulle, $\sum_{j=1}^n g_j \cos(\omega j)$ et $\sum_{j=1}^n g_j \sin(\omega j)$ sont proches de zéro. De plus, on obtient pour la variance

$$\begin{aligned} \frac{1}{n^4} \sum_{j,k,l,m=1}^n \mathbb{E} [\varepsilon_j \varepsilon_k \varepsilon_l \varepsilon_m] \cos(\omega(j-k)) \cos(\omega(l-m)) - \frac{\sigma^4}{n^2} &= 2 \frac{\sigma^4}{n^2} - \frac{\sigma^4}{n^2} \\ &= \frac{\sigma^4}{n^2}. \end{aligned}$$

Cela montre que même si les données contiennent un signal périodique de fréquence ω_0 , la distribution du périodogramme pour les fréquences de Fourier différentes de la plus proche de ω_0 , le périodogramme se comporte comme s'il n'y avait que du bruit. Cela est d'autant plus vrai lorsque n est grand. Dans le cas de données irrégulières, le périodogramme perd sa propriété d'invariance par translation (dans le temps) et d'orthogonalité des fonctions trigonométriques $\cos(\omega t)$ et $\sin(\omega t)$ qui composent les fonctions trigonométriques complexes $\exp(-i\omega t)$. Lomb [17] propose une modification du périodogramme qui permet de retrouver la propriété d'invariance par translation et d'orthogonalité des fonctions trigonométriques $\cos(\omega t)$ et $\sin(\omega t)$. Scargle [26] étudie le comportement de ce périodogramme modifié. Cette approche porte donc le nom de la méthode de Lomb-Scargle. De plus, Scargle montre l'équivalence entre le périodogramme modifié et l'ajustement par la méthode des moindres carrés d'une fonction de

régression de la forme $A \cos(\omega t) + B \sin(\omega t)$, $A, B \in \mathbb{R}$. Ici la constante μ n'est pas présente dans le modèle parce qu'on l'ajuste aux données centrées, c'est-à-dire que l'on soustrait la moyenne \bar{y} aux données y_1, \dots, y_n . Dans [35], les auteurs proposent la méthode "floating mean periodogram" qui est la méthode de Lomb-Scargle avec le paramètre μ dans le modèle. Ils montrent d'ailleurs que cette méthode est équivalente à la méthode des moindres carrés avec une fonction de régression $A \cos(\omega t) + B \sin(\omega t) + \mu$, $A, B, \mu \in \mathbb{R}$.

1.8 L'autocorrélation

Comme définie dans la section (1.6), l'autocovariance pour un processus $Y(t)$ faiblement stationnaire, s'écrit

$$\gamma(s) = \mathbb{E}[Y(t)Y(t+s)] - \mu^2. \quad (1.11)$$

L'autocorrélation est donc simplement

$$\rho(s) = \frac{\gamma(s)}{\sigma^2}, \quad (1.12)$$

où $\sigma^2 = \text{Var}[Y]$. Pour un signal périodique $g(t)$, l'autocorrélation permet d'estimer la période. En effet, si le décalage est égal à la période ou à un multiple de la période, l'autocorrélation est proche de 1 ($-1 \leq \rho(s) \leq 1$). Tandis que si le décalage est égal à la demi-période ou à un multiple de la demi-période, alors l'autocorrélation est proche de -1 . Dans le cas d'un échantillon de données avec espacement régulier de valeur 1, les décalages possibles s ne sont plus continus mais un ensemble fini $\mathcal{D} = \{1, \dots, (n-1)\}$. On retrouve $n-1$ décalages de longueur 1, $n-2$ décalages de longueur 2, ... et 1 décalage de longueur $(n-1)$. Une estimation $\hat{\rho}(k)$ de l'autocorrélation $\rho(k)$ est

$$\hat{\rho}(k) = \frac{1}{n-1} \frac{\sum_{i=1}^{n-k} (y_i - \bar{y})(y_{i+k} - \bar{y})}{s^2}, \quad (1.13)$$

avec $y_i = y(i)$ et $s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$. Reprenons le signal périodique (1.10). La figure 1.6 illustre l'autocorrélation (1.13). Dans le cas d'observations irrégulièrement espacées, le nombre de couples avec un décalage fixe ne vérifie plus la propriété de décroissance en fonction de la valeur du décalage. La figure 1.7 représente l'histogramme des décalages pour les temps irréguliers d'observations de l'étoile HIP 8, figure 1.3. En 1986, Gaskell et Sparke [12] introduisent une méthode d'interpolation de l'autocorrélation, modifiée par Gaskell et Peterson [11] en 1987, pour contourner le problème des trous apparaissant dans la fonction d'autocorrélation, dus à l'échantillonnage irrégulier. En 1988, Edelson et Krolik [6] proposent de créer des bins de décalages. Les bins $\mathcal{B}(k)$ sont tous de même longueur. Une estimation des coefficients, l'estimateur "Discrete Correlation Function" DCF(k), pour chaque bin est proposée.

$$\text{DCF}(k) = \frac{1}{M} \sum_{i,j \in \mathcal{B}(k)} \frac{(y_i - \bar{y})(y_j - \bar{y})}{\sigma^2}, \quad (1.14)$$

où

$$\mathcal{B}(k) = \left\{ i, j \mid k\Delta - \frac{\Delta}{2} < |t_i - t_j| < k\Delta + \frac{\Delta}{2} \right\},$$

et $M = \text{card}(\mathcal{B})$. La figure 1.8 représente l'estimateur DCF (1.14) des données de l'étoile HIP 8, figure 1.3.

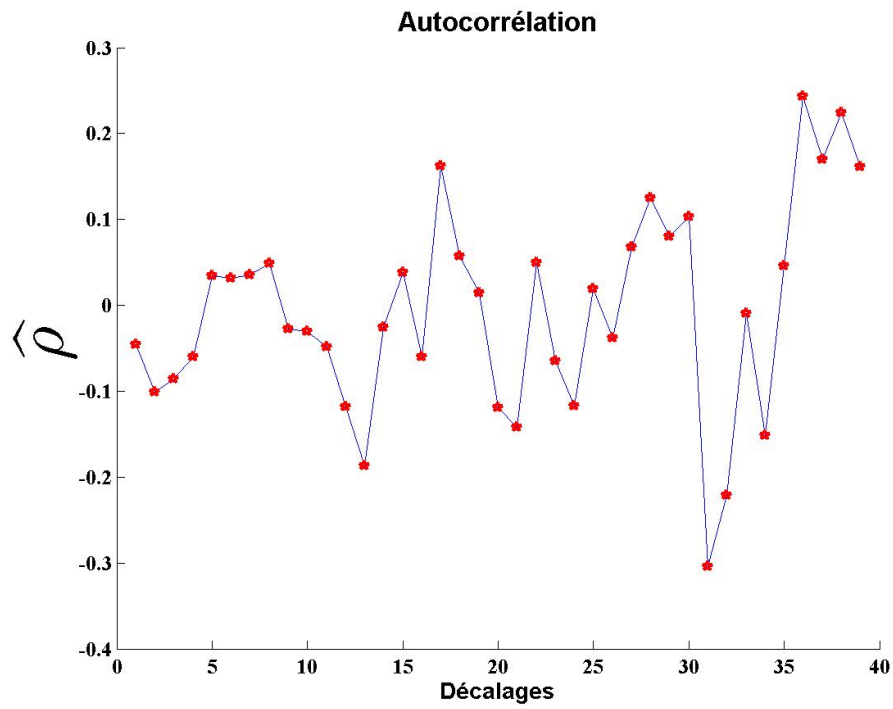


FIG. 1.6 – Autocorrélation estimée (1.13) des données pour le signal périodique (1.10).

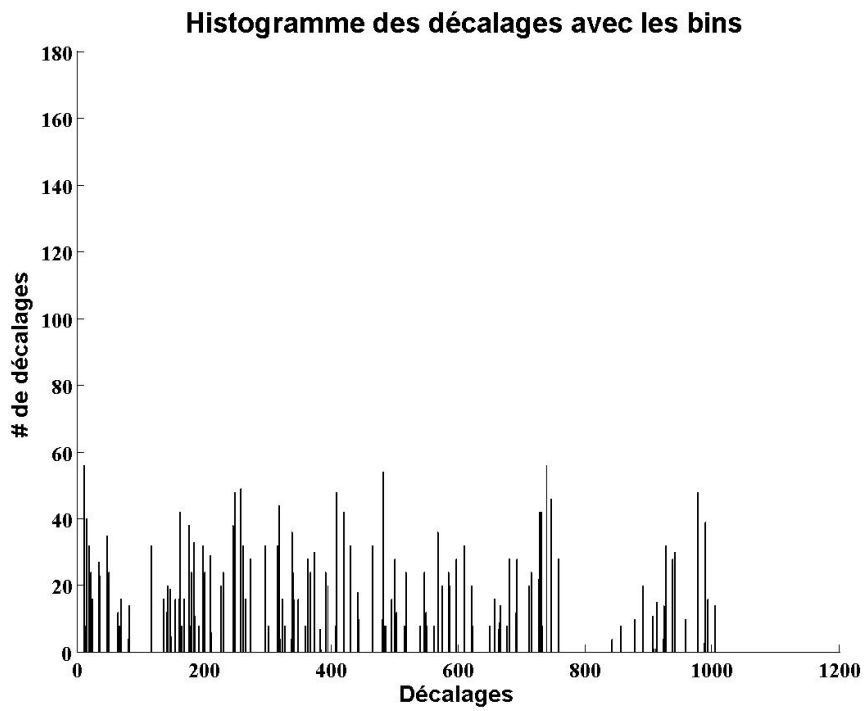


FIG. 1.7 – Histogramme des décalages pour les temps d’observations de l’étoile HIP 8 de type Mira, figure 1.3.

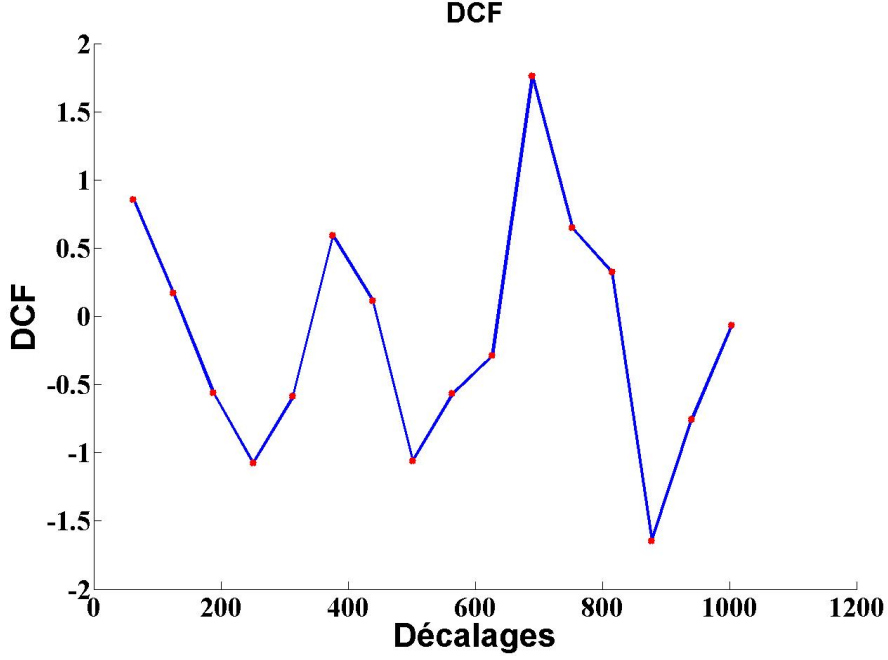


FIG. 1.8 – L’estimateur DCF (1.14) des données de l’étoile HIP 8 de type Mira, figure 1.3.

Ici, nous considérons que le signal sous-jacent est déterministe, contrairement aux auteurs qui supposent le signal comme étant stochastique. Ainsi pour nous, σ^2 est uniquement la variance du processus stochastique représentant les erreurs. Nous avons testé cette méthode sur les données Hipparcos, mais sans succès. Nous avons essayé d’autres méthodes de binning, comme par exemple en fonction de la densité locale des décalages, ou encore en fixant un nombre constant de décalages dans chaque bin, cela n’a rien donné non plus. Écrivons $\hat{\rho}(v)$, $v = 1, \dots, M$, avec $M \leq \frac{n(n-1)}{2}$, les estimations des coefficients de corrélation $\hat{\rho}(|t_i - t_j|) = \frac{\hat{\gamma}(|t_i - t_j|)}{\sigma^2}$. La méthode qui a donné de meilleurs résultats est basée sur un lissage de la fonction d’autocorrélation estimée $\hat{\rho}(v)$ correspondant au v -ème décalage,

$$\rho^{lisse}(k - 1 + \lfloor \frac{m}{2} \rfloor) = \frac{1}{m} \sum_{v=k}^{k+m} \hat{\rho}(v), \quad (1.15)$$

avec $m > 0$, la taille du lisseur, à déterminer en fonction de la densité des décalages, par exemple, et $k \geq 1$. De plus $\rho^{lisse}(l) = \rho^{lisse}(\lfloor \frac{m}{2} \rfloor)$, $l < \lfloor \frac{m}{2} \rfloor$ et $\rho^{lisse}(l) = \rho^{lisse}(\frac{n(n-1)}{2} - \lfloor \frac{m}{2} \rfloor)$, $l > \frac{n(n-1)}{2} - \lfloor \frac{m}{2} \rfloor$. Lors d’un échantillonnage irrégulier engendrant peu de décalages non observés, cette méthode fonctionne. Lorsque l’échantillonnage irrégulier engendre de grands nombres de décalages inobservés, cette méthode ne fonctionne que moyennement pour des signaux avec des périodes assez grandes, et pas du tout, pour des signaux avec des petites périodes. Dans le catalogue Hipparcos, les étoiles ayant été caractérisées par cette méthode ont des périodes de variations de leur magnitude comprises entre 100 et 400 jours. L’échantillonnage Hipparcos étant plus du type irrégulier avec de grands nombres de décalages inobservés, cette méthode n’a pas donné de bons résultats. Au chapitre 2 sur les données Hipparcos, la performance de cette méthode est comparée avec les performances des méthodes de l’analyse harmonique, du périodogramme en amplitude (Deeming) et de la méthode non paramétrique.

1.9 Théorie de Fourier

Dans cette section, on introduit la théorie de Fourier, le fondement de l’analyse spectrale d’une fonction $f(t)$ quelconque donnée. On donnera aussi les hypothèses minimales que la fonction $f(t)$

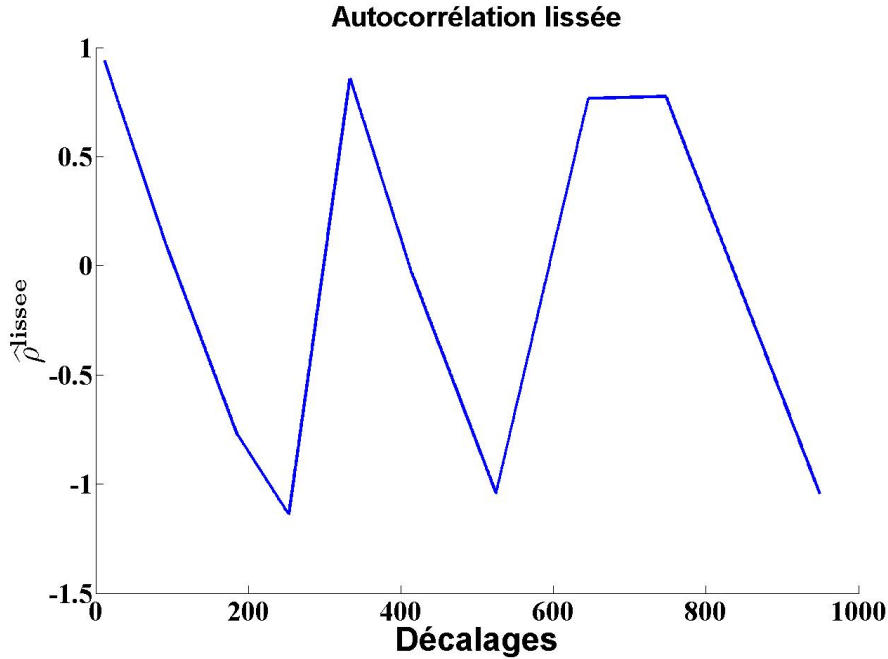


FIG. 1.9 – L'autocorrélation lissée (1.15) des données de l'étoile HIP 8 de type Mira, figure 1.3.

doit vérifier pour pouvoir utiliser la théorie de Fourier. Nous considérons les fonctions f dans l'espace $L^2(a, b)$ des fonctions de carré intégrable sur $[a, b]$, $b > a$. Il s'agit d'un espace de Hilbert, qui est un espace de Banach complet dont la norme $\|f\| = \sqrt{\langle f, f \rangle}$ provient du produit scalaire $\langle f, g \rangle = \int_a^b f(t)\bar{g}(t)dt$, $f, g \in \mathbb{C}$, où \bar{g} est le conjugué complexe de g . Un autre espace fonctionnel important est l'espace $L^1(a, b)$ des fonctions intégrables sur $[a, b]$. Ce n'est pas un espace de Hilbert, mais il contient l'espace $L^2(a, b)$. Dans la partie suivante, on considère des fonctions périodiques $f(t)$ de période 2π . La période n'est pas restrictive. En effet, on peut écrire une fonction de période P en considérant le changement de variable $h(t) = P\frac{t}{2\pi}$.

1.9.1 Théorie des séries de Fourier

Soit $f : \mathbb{R} \rightarrow \mathbb{C}$ une fonction 2π -périodique. Nous allons donner les propriétés que la fonction $f(x)$ doit vérifier pour que le développement de $f(x)$ en une série infinie (1.16) soit possible.

$$f(x) = \sum_{n=-\infty}^{\infty} c_n \exp(inx), \quad (1.16)$$

où

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \exp(-inx) dx. \quad (1.17)$$

On peut aussi l'écrire sous la forme d'une série infinie réelle

$$f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos(nx) + b_n \sin(nx), \quad (1.18)$$

où

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(nx) dx, \quad (1.19)$$

et

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(nx) dx. \quad (1.20)$$

En effet, pour $k \geq 1$, on a $a_k = c_{-k} + c_k$, et $b_k = i(c_k - c_{-k})$ car

$$\begin{aligned} c_{-k} + c_k &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) (\exp(inx) + \exp(-inx)) dx = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) 2 \cos(nx) dx, \\ &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(nx) dx, \end{aligned} \quad (1.21)$$

et

$$\begin{aligned} i(c_k - c_{-k}) &= \frac{i}{2\pi} \int_{-\pi}^{\pi} f(x) (\exp(-inx) - \exp(inx)) dx = \frac{i}{2\pi} \int_{-\pi}^{\pi} f(x) (-2i) \sin(nx) dx \\ &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(nx) dx, \end{aligned} \quad (1.22)$$

pour $k = 0$, on a $a_0 = 2c_0$, car

$$\begin{aligned} 2c_0 &= c_0 + c_0 \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \exp(i0x) dx = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) (\exp(i0x) + \exp(i0x)) dx \\ &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(0x) dx = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) dx. \end{aligned} \quad (1.23)$$

Si la fonction f est intégrable dans l'intervalle $[-\pi, \pi]$, c'est-à-dire

$$\int_{-\pi}^{\pi} |f(x)| dx < \infty, \quad (1.24)$$

alors les coefficients (1.17), (1.19) et (1.20) sont bien définis. Maintenant introduisons les sommes partielles d'une série de Fourier. Cela nous permettra de définir le noyau de Dirichlet et de Féjer.

1.9.2 Sommes partielles d'une série de Fourier

Soit $f : \mathbb{R} \rightarrow \mathbb{C}$ une fonction 2π -périodique, $f \in L^1([-\pi, \pi])$ (f est dit intégrable). Définissons la n -ème somme partielle de la série de Fourier complexe de f par

$$S_n(f; x) = \sum_{k=-n}^n c_k \exp(ikx), \quad n = 0, 1, \dots, x \in \mathbb{R}. \quad (1.25)$$

On peut aussi définir la moyenne des $n + 1$ premières sommes partielles

$$\sigma_n(f; x) = \frac{1}{n+1} \sum_{k=0}^n S_k(f; x), \quad n = 0, 1, \dots, x \in \mathbb{R}. \quad (1.26)$$

On appelle $\sigma_n(f; x)$ la moyenne de Cesàro. Ainsi, on peut écrire $S_n(f; x)$ et $\sigma_n(f; x)$, comme

$$S_n(f; x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) D_n(x-t) dt \quad (1.27)$$

$$\sigma_n(f; x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) K_n(x-t) dt, \quad (1.28)$$

avec

$$D_n(x) = \sum_{k=-n}^n \exp(ikx) = \begin{cases} \frac{\sin((n+\frac{1}{2})x)}{\sin(\frac{x}{2})} & \text{si } x \notin 2\pi\mathbb{Z} \\ 2n+1 & \text{si } x \in 2\pi\mathbb{Z} \end{cases}, \quad (1.29)$$

le noyau de Dirichlet, et

$$K_n(x) = \frac{1}{n+1} \sum_{k=0}^n D_k(x) = \begin{cases} \frac{1}{n+1} \left[\frac{\sin(\frac{n+1}{2}x)}{\sin(\frac{x}{2})} \right]^2 & \text{si } x \notin 2\pi\mathbb{Z} \\ 1 & \text{si } x \in 2\pi\mathbb{Z} \end{cases},$$

le noyau de Féjer. En effet,

$$\begin{aligned} S_n(f; x) &= \sum_{k=-n}^n c_k \exp(ikx) = \frac{1}{2\pi} \sum_{k=-n}^n \int_{-\pi}^{\pi} f(t) \exp(-ikt) dt \exp(ikx) \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \sum_{k=-n}^n \exp(ik(x-t)) dt. \end{aligned} \quad (1.30)$$

$$\begin{aligned} \sigma_n(f; x) &= \frac{1}{n+1} \sum_{k=0}^n S_k(f; x) = \frac{1}{2\pi} \frac{1}{n+1} \sum_{k=0}^n \int_{-\pi}^{\pi} f(t) D_k(x-t) dt \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \frac{1}{n+1} \sum_{k=0}^n D_k(x-t) dt. \end{aligned} \quad (1.31)$$

Le théorème de Fejér justifie le fait que toute fonction 2π -périodique et continue sur l'intervalle $[-\pi, \pi]$ peut être développée dans une série de Fourier. Les résultats qui en découlent montrent que même des fonctions seulement intégrables sur l'intervalle $[-\pi, \pi]$ peuvent aussi être développées en série de Fourier.

Théorème 1 * [Fejér] Soit $f : \mathbb{R} \rightarrow \mathbb{C}$ une fonction 2π -périodique continue et bornée sur l'intervalle $[-\pi, \pi]$. Alors

$$\sup_{x \in \mathbb{R}} |\sigma_n(f; x) - f(x)| \xrightarrow{n \rightarrow \infty} 0, \quad (1.32)$$

où $\sigma_n(f; x)$ est la moyenne de Cesàro définie en (1.26).

Supposons maintenant que $f : \mathbb{R} \rightarrow \mathbb{C}$ est seulement intégrable sur $[-\pi, \pi]$. Un théorème de Lebesgue dit que

Théorème 2 * [de Lebesgue] Soit $f : \mathbb{R} \rightarrow \mathbb{C}$ une fonction intégrable sur $[-\pi, \pi]$. Alors

$$\lim_{n \rightarrow \infty} \sigma_n(f; x) = f(x), \text{ p.p. } x \in \mathbb{R}. \quad (1.33)$$

Le théorème de Fejér donne la condition que f soit 2π -périodique, intégrable et continue pour que les sommes partielles $\{S_k\}_{k \geq 0}$ de la série de Fourier de $f(x)$ converge uniformément au sens de Cesàro vers $f(x)$, $\forall x \in \mathbb{R}$. Le théorème de Lebesgue, moins restrictif que le théorème de Fejér, dit que si f est 2π -périodique et intégrable sur $[-\pi, \pi]$, alors la série de Fourier converge vers $f(x)$, pour tout $x \in \mathbb{R}$, en dehors d'un ensemble de points au plus dénombrable (convergence presque partout). A toute fonction f 2π -périodique de carré intégrable ($f \in L^2([-\pi, \pi])$) sur $[-\pi, \pi]$, on peut associer une série de Fourier complexe (1.16). En effet, les fonctions trigonométriques $\{\exp(ikx)\}_{k \in \mathbb{Z}}$ forment une base orthogonale de l'espace $L^2([-\pi, \pi])$, et donc $f(x) = \sum_{k=-\infty}^{\infty} c_k \exp(ikx)$, où les c_k sont les coefficients de Fourier.

1.9.3 La transformée de Fourier

Soit $f \in L^1([-\pi, \pi])$. La transformée de Fourier est définie comme

$$\mathcal{TF}\{f\}(\omega) = \int_{-\infty}^{+\infty} f(t)e^{-i\omega t} dt \in \mathbb{C}. \quad (1.34)$$

Si $\mathcal{TF}\{f\} \in L^1(\mathbb{R})$, la transformée de Fourier inverse est

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \mathcal{TF}\{f\}(\omega)e^{i\omega t} d\omega. \quad (1.35)$$

Une propriété importante de la transformée est que l'énergie du signal est conservée. Ceci transparait dans le théorème de Parseval.

Théorème 3 (Parseval) $\forall f, g \in L^2(\mathbb{R})$.

$$\langle f, g \rangle = \int_{-\infty}^{+\infty} f(t)\bar{g}(t)dt = \int_{-\infty}^{\infty} \mathcal{TF}\{f\}(\omega)\overline{\mathcal{TF}\{g\}(\omega)}d\omega = \langle \mathcal{TF}\{f\}, \mathcal{TF}\{g\} \rangle .$$

Un important résultat est le lemme de Riemann-Lesbesgue :

Lemme 1 (Riemann-Lesbesgue) Soit $f : [a, b] \rightarrow \mathbb{C}$ une fonction intégrable où $[a, b]$ est un intervalle compact de \mathbb{R} . Alors $\lim_{|\omega| \rightarrow \infty} \int_a^b f(x) \exp(i\omega x) dx = 0$, où $\omega \in \mathbb{R}$.

Un résultat immédiat du lemme 1 est que si $f : \mathbb{R} \rightarrow \mathbb{C}$ est 2π -périodique et intégrable dans $[-\pi, \pi]$, alors $\lim_{|n| \rightarrow \infty} c_n(f) = 0$ où $c_n(f)$ sont les coefficients de Fourier complexes (qui sont les transformées de Fourier) de f .

1.10 Echantillonnage

L'échantillonnage est la première étape indispensable à toute analyse statistique. En effet, l'ordinateur ne peut pas traiter les signaux analogiques ou continus. Il ne traite que des ensembles numériques. Il existe une infinité d'échantillonnages différents. Un processus d'échantillonnage, dans le cas dit régulier, est défini par sa fréquence d'échantillonnage $\nu_e = \frac{1}{\delta_e}$, ou $\omega_e = \frac{2\pi}{\delta_e}$, le nombre d'observations par unité de temps, où δ_e est le pas d'échantillonnage. La fréquence ν_e et l'espacement δ_e sont constants, contrairement au cas irrégulier, où ν_e et δ_e peuvent prendre plusieurs valeurs allant d'un nombre fini à un nombre infini de valeurs, avec une fonction de probabilité associée. L'échantillonnage produit une série temporelle de couples temps/observations (t_i, y_i) , $i = 1, \dots, n$. En terme de coût de mémoire et de temps de calculs, il est préférable d'avoir n pas trop grand. Cependant, on ne peut pas choisir ν_e aussi petit que l'on veut. Par exemple, dans le cas régulier, ν_e doit vérifier la condition d'être au minimum deux fois plus élevée que la plus haute fréquence présente dans le spectre du signal. Cette fréquence est nommée la fréquence de Nyquist ν_{Nyquist} .

1.10.1 Echantillonnage régulier

Ici, nous considérons des temps d'observations réguliers $t_i = i\delta_e$, avec $\delta_e > 0$ le pas de discrétisation. Pour pouvoir recouvrir un signal périodique $g(t) = \sin(\omega_0 t)$, de fréquence ω_0 , à partir d'un ensemble de n observations du signal $\{Y(t_1), \dots, Y(t_n)\}$, sans bruit, avec les t_i comme ci-dessus, la fréquence d'échantillonnage ω_e doit vérifier la condition $\omega_0 \leq \frac{\omega_e}{2}$. En effet, il existe une infinité de fonctions sinusoidales de fréquences $\{\omega_{0,k}\}_{k=1}^{\infty}$, qui passent par les points $\{Y(t_1), \dots, Y(t_n)\}$. Les fréquences $\omega_{0,k}$ portent le nom d'alias de ω_0 . Cela est dû au fait que, avec une fréquence d'échantillonnage ω_e , la fréquence maximale détectable est $\frac{\omega_e}{2}$. Maintenant, considérons une fréquence $\omega_{e_1} < 2\omega_0$. Alors, il y a moins de trois observations par période. Avec un tel échantillon, on peut trouver plusieurs fonctions sinusoidales avec chacune une fréquence différente passant exactement par toutes les observations. Dans le cas contraire, $\nu_{e_1} \geq 2\omega_0$, il y a au moins 3 observations par période. Ainsi, en connaissant la forme du signal $g(t)$, il est possible de le recouvrir avec la bonne fréquence. Finalement, considérons un signal sinusoidal pour illustrer la problématique des alias. Posons $\omega_{\text{Nyquist}} = 2\pi\nu_{\text{Nyquist}} > 0$ la fréquence maximale détectable, $\delta \in (0, \nu_{\text{Nyquist}})$ et $t = \frac{i}{\nu_e}$, pour un $i \in \{0, 1, \dots, n-1\}$. Donc,

$$\sin(2\pi(\nu_{\text{Nyquist}} + \delta)t) = \sin(2\pi(-\nu_{\text{Nyquist}} + \delta)t) \quad (1.36)$$

$$\iff$$

$$\sin(2\pi\nu_{\text{Nyquist}}t + 2\pi\delta t) = \sin(-2\pi\nu_{\text{Nyquist}}t + 2\pi\delta t). \quad (1.37)$$

A l'équation (1.36), on utilise la définition de la fréquence de ω_{Nyquist} . Le résultat de l'équation (1.37) implique que $2\pi\nu_{\text{Nyquist}}t + 2\pi\delta t = -2\pi\nu_{\text{Nyquist}}t + 2\pi\delta t + 2\pi\nu_e t$. En effet, avec des temps $t = \frac{i}{\nu_e}$, $2\pi\nu_e t = 2i\pi$. On obtient que $4\pi \frac{\nu_{\text{Nyquist}}}{\nu_e} i = 2\pi \frac{i}{\nu_e}$, ce qui est équivalent à $\nu_{\text{Nyquist}} = \frac{\nu_e}{2}$. La fréquence $\nu_{\text{Nyquist}} = \frac{\nu_e}{2}$ est la fréquence de Nyquist. C'est pourquoi, lors d'un échantillonnage de fréquence ν_e , on se limite aux fréquences de l'intervalle $[0, \frac{\nu_e}{2}]$ ou $[\frac{\nu_e}{2}, \nu_e]$ ou $[\nu_e, \frac{3\nu_e}{2}]$ ou Un signal de fréquence $\omega_0 > 0$ échantillonné régulièrement avec une fréquence d'échantillonnage ω_e , peut être reconstruit à partir des observations si et seulement si $\omega_0 \leq \frac{\omega_e}{2}$. Le même raisonnement est valable pour tout signal $g(t)$ périodique quelconque. En effet, étant donné que toute fonction périodique peut s'écrire comme une série de Fourier, la fréquence de Nyquist devient le double de la fréquence maximale du spectre du signal.

1.10.2 Echantillonnage irrégulier

Comme son nom l'indique, l'échantillonnage irrégulier implique que la fréquence d'échantillonnage ν_e n'est plus constante au cours du temps. Ainsi la notion de fréquence de Nyquist n'est plus aussi claire. En effet, la fréquence de Nyquist apparaît à cause de la périodicité dans l'échantillonnage. Nous allons distinguer deux principaux modèles d'irrégularité de l'échantillonnage.

Irreg₁ L'échantillonnage à données manquantes. Le signal est échantillonné avec une fréquence ν_e constante, mais un nombre de données est manquant. Jones [14] est le premier à proposer un tel modèle de données manquantes dans le cadre de l'estimation spectrale d'un processus stochastique stationnaire, voir la section 1.11 sur l'historique de la recherche des fréquences inconnues. Ce modèle d'échantillonnage engendre une fréquence de Nyquist.

Irreg₂ L'échantillonnage est un processus aléatoire.

Le modèle d'échantillonnage Irreg₁ étant déjà bien connu, on va se concentrer sur le modèle Irreg₂. En premier lieu, définissons le repliement en phase des observations.

Définition 1 (Repliement en phase) Soit un signal $g(t)$ de période $P_0 = \frac{2\pi}{\omega_0}$ et soit le signal échantillonné g_1, \dots, g_n aux temps t_1, \dots, t_n . Un repliement en phase avec la période P_0 est la translation temporelle de tous les temps t_1, \dots, t_n d'observations dans l'intervalle $[0, P_0)$. Les temps translatés $\tilde{t}_1, \dots, \tilde{t}_n$ dans l'intervalle $[0, P_0)$ sont tels que $\tilde{t}_i = t_i - kP_0$, où $k \in \mathbb{N}$ vérifie la condition $t_i - (k+1)P_0 < 0$. Soit $\tilde{t}_{(1)}, \dots, \tilde{t}_{(n)}$ les valeurs ordonnées de $\tilde{t}_1, \dots, \tilde{t}_n$.

Nous allons séparer les processus aléatoires d'échantillonnage en deux groupes. Pour cela, nous introduisons la notion de couverture de l'intervalle de la période du signal.

Définition 2 (Couverture asymptotique de la période) Soit un signal $g(t)$ de période $P_0 = \frac{2\pi}{\omega_0}$ et soit le signal échantillonné g_1, \dots, g_n aux temps t_1, \dots, t_n . Maintenant considérons les temps ordonnés $\tilde{t}_{(1)}, \dots, \tilde{t}_{(n)}$, après un repliement en phase. Alors, lorsque $n \rightarrow \infty$, $\max_{i=1, \dots, n-1} |\tilde{t}_{(i+1)} - \tilde{t}_{(i)}| \rightarrow 0$, $\tilde{t}_{(1)} \rightarrow 0$, $\tilde{t}_{(n)} \rightarrow P_0$. Ainsi $\{\tilde{t}_{(i)}\}_{i=1}^n$ recouvre tout l'intervalle $[0, P_0)$.

1. Les processus d'échantillonnages aléatoires, vérifiant la définition (2), permettent, asymptotiquement, une couverture uniforme de l'intervalle $[0, P_0)$. Dans [13], les auteurs définissent la i -ème observation provenant d'un échantillonnage uniforme de n observations comme étant $nT_{(i)}$, où $T_{(i)}$ est la i -ème statistique d'ordre de l'ensemble de variables aléatoires $\{T_i\}_{i=1}^n$ avec $T_i \sim \mathcal{U}(0, 1)$. Le facteur n veut dire que l'espacement entre deux temps successifs est environ 1. Si un espacement δ est souhaité, alors le facteur $n\delta$ doit être utilisé. Donnons deux autres exemples de processus entrant dans cette catégorie.
 - (a) Les processus de renouvellement, où la i -ème observation $T_i = \sum_{j=1}^i V_j$, où $V_j \stackrel{\text{iid}}{\sim} F$, distribuées selon une loi F , est une variable aléatoire strictement positive. Un exemple classique est le processus de Poisson, où $F = \mathcal{E}(\lambda)$, avec $\lambda > 0$. Ici l'espacement moyen est $\frac{1}{\lambda}$. On peut démontrer que cette séquence couvre asymptotiquement chaque période P_0 . Ceci sera discuté plus loin.
 - (b) Les processus perturbés (jittering, en anglais), où la i -ème observation $T_i = i + U_i$ où $U_i \sim \mathcal{U}(-\frac{1}{2}, \frac{1}{2})$. Ici l'espacement moyen est 1. Cet échantillonnage couvre également toutes les périodes $P_0 > 0$.
2. Des processus d'échantillonnages aléatoires ne vérifiant pas la condition (2) sont par exemple des processus dont le modèle d'échantillonnage est en phase avec le signal périodique de période P_0 . Par exemple, posons $T_i = i\delta + U_i$ où $U_i \sim \mathcal{U}(0, \beta P_0)$, $\delta = P_0$ et $1 > \beta > 0$. Lorsque $n \rightarrow \infty$, un repliement en phase ne donne qu'une couverture partielle de l'intervalle $[0, P_0)$.

Dans le chapitre 5 sur les simulations, plusieurs modèles d'échantillonnage seront présentés. Un très grand nombre de modèles d'échantillonnage sont des modèles dit additifs (ou aussi appelé temps des occurrences d'un processus de renouvellement). On les écrit

$$T_n = \sum_{i=1}^n V_i,$$

où $V_i \sim F$, $i = 1, \dots, n$ avec fonction de densité f identiquement nulle pour les valeurs négatives ou nulles. Shapiro et Silverman [29] donnent des conditions sur la fonction de densité f de la loi de V_i , pour que, lorsque $n \rightarrow \infty$, il n'y ait plus de problème d'alias, contrairement, au cas de l'échantillonnage régulier ou à données manquantes, par exemple. Un tel processus vérifiant ces conditions est dit sans alias. Cela se traduit par une fréquence de Nyquist infinie. Un tel processus couvre asymptotiquement chaque période $P_0 > 0$. Un exemple de processus sans alias est le processus de Poisson. En effet, la condition qu'ils posent pour que le processus additif soit sans alias est la suivante. La fonction caractéristique (conjugué complexe de la transformée de Fourier de la fonction de densité f)

$$\phi(\omega) = \int_0^{+\infty} f(t) \exp(i\omega t) dt \quad (1.38)$$

doit être injective pour $\omega \in \mathbb{R}$. La fonction caractéristique d'une variable aléatoire exponentielle $\mathcal{E}(\lambda)$ vaut

$$\phi(\omega) = \int_0^{+\infty} \lambda \exp(-\lambda t) \exp(i\omega t) dt = \frac{\lambda}{\lambda - i\omega}. \quad (1.39)$$

Clairement, pour $\omega_1 \neq \omega_2$, $\phi(\omega_1) \neq \phi(\omega_2)$. Cette condition donne aussi une condition sur le comportement de la fonction de densité f , comme Shapiro et Silverman le montrent dans [29]. Ils montrent que la fonction de densité f doit être strictement décroissante et que la transformée en sinus $\int_0^{+\infty} f(t) \sin(\omega t) dt$ doit être strictement positive. Un processus ne vérifiant pas ces conditions engendre l'existence de la fréquence de Nyquist. Dans ce cas, on peut avoir besoin de la connaître. On peut se référer à Eyer et Bartholdi [7]. Ils donnent une méthode d'estimation de cette fréquence limite. De plus, ils montrent que lors d'échantillonnage irrégulier, la fréquence de Nyquist est en général très élevée. Mignard [23], dans une "technical note", fournit une condition basée sur la périodicité de la fenêtre spectrale qui généralise la fréquence de Nyquist. Dans le cas de l'échantillonnage d'Hipparcos, qui est un échantillonnage irrégulier contenant une composante régulière avec une fréquence d'échantillonnage élevée (voir section 1.2 de l'introduction), la fréquence de Nyquist est très élevée. Cela permet de détecter des hautes fréquences.

1.11 La recherche de fréquence dans la littérature scientifique

La recherche de l'existence d'une période dans une série temporelle et sa caractérisation est un problème qui a été investigué à la fin du 19ème siècle, pour des données régulièrement espacées. On trouve la première définition du périodogramme

$$R = \sqrt{A^2 + B^2}, \quad (1.40)$$

où $A \frac{T}{2} = \int_{t_1}^{t_1+T} g(t) \cos(k\omega t) dt$ et $B \frac{T}{2} = \int_{t_1}^{t_1+T} g(t) \sin(k\omega t) dt$, avec $\omega = \frac{2\pi}{T}$, $k \in \mathcal{N}$, dans l'article de Schuster [27], en 1898. Plus tard, en 1929, Fisher [9] propose un test pour la détection de la période sous-jacente. Il compare le rapport de la plus grande valeur du périodogramme et de la valeur moyenne du périodogramme à un certain seuil de détection. Ce test est connu sous le nom de test de Fisher (on peut voir, par exemple, [3], p. 339). Les scientifiques ont commencé par s'intéresser à des processus de temps d'observations irréguliers. Dans bien des domaines, comme par exemple l'astronomie, la météorologie, la sismologie, les données irrégulièrement espacées sont courantes. Par exemple, en astronomie, lorsque les observations sont faites depuis la terre, la première périodicité apparaissant dans l'échantillonnage est celle de 1 jour, correspondant aux observations faites chaque nuit. Ensuite, une deuxième périodicité apparaissant dans l'échantillonnage provient de la fréquence des observations effectuées pendant la nuit. Nous avons ainsi une suite d'observations avec des espacements d'environ 12 heures entre la fin et le début de chaque groupe d'observations et des espacements de par exemple 1 minute entre chaque observations effectuées pendant la nuit. La couverture nuageuse pouvant perturber les observations, l'échantillonnage peut contenir des espacements entre les observations pouvant être plus grands que 12 heures pour les espacements entre les groupes et plus grands que 1 minute pour les espacements dans le groupe. Plus généralement, dans ce type d'échantillonnage, la fréquence d'échantillonnage est presque constante, mais occasionnellement, une ou plusieurs observations sont manquantes. C'est par ce type de processus, dit à données manquantes, que les premières investigations ont débuté. Jones [14], en 1962, étudie la variance asymptotique de l'estimation de la densité spectrale, dans le cas spécifique où des groupes d'observations sont manquantes périodiquement. Il considère l'estimateur usuel de la densité spectrale $f_{sp}(\omega)$, pour $n \rightarrow \infty$,

$$\hat{f}_{sp}(\omega) = \frac{1}{n} \sum_{j=1}^n \sum_{l=1}^n (y_j - \bar{y})(y_l - \bar{y}) w_{j-l}^n \cos(\omega(j-l)), \quad (1.41)$$

où $\sum_{k=0}^n \omega_k^n \cos(k\omega)$ est la fenêtre spectrale. Il trouve que la variance de (1.41) est la somme de la variance de l'estimateur dans le cas d'échantillonnage régulier et de termes formés des fréquences correspondantes à la périodicité des observations manquées. En 1969, Bloomfield [1] propose un estimateur de la densité spectrale dans le cas d'un processus à données manquantes plus général. Barning

est le premier à s'intéresser à une approche par moindres carrés pour la détection de fréquences inconnues, avec échantillonnage irrégulier, en 1963, en ajustant par moindres carrés une fonction de régression sinusoidale à une fréquence donnée, puis en retranchant la fonction ajustée aux données. Il répète cette procédure, avec une fréquence différente donnée, jusqu'à ce que la somme des carrés des résidus soit plus petite qu'un certain seuil. Les fréquences considérées sont déterminées par le périodogramme. Cette méthode porte actuellement le nom de "matching pursuit". Vaniček [30] reprend cette approche en 1969. Il propose de minimiser la distance quadratique moyenne

$$\sum_{i=1}^n (y_i - h_i)^2$$

entre les observations y_i régulièrement espacées et une fonction h_i formée de plusieurs composantes sinusoidales. En 1970, Vaniček [31] applique cette méthode à des données irrégulièrement espacées dans le temps sans que le processus d'échantillonnage ne vérifie un quelconque processus à données manquantes. En 1975, Deeming [5] propose de considérer une estimation du périodogramme en amplitude

$$P(\omega) = \frac{2}{n} \sqrt{\left(\sum_{i=1}^n y_i \cos(\omega t_i)\right)^2 + \left(\sum_{i=1}^n y_i \sin(\omega t_i)\right)^2}, \quad (1.42)$$

comme si les fréquences de Fourier indépendantes étaient au nombre de $\frac{n}{2}$, si n est paire, $\lceil \frac{n-1}{2} \rceil$, si n est impair. Lomb [17], en 1976, simplifie la méthode de Vaniček [31] pour donner le périodogramme modifié,

$$\frac{(\sum_{i=1}^n y_i \cos(\omega(t_i - \tau)))^2}{\sum_{i=1}^n \cos(\omega(t_i - \tau))^2} + \frac{(\sum_{i=1}^n y_i \sin(\omega(t_i - \tau)))^2}{\sum_{i=1}^n \sin(\omega(t_i - \tau))^2}, \quad (1.43)$$

applicable aux données irrégulières. Ce périodogramme modifié vérifie la propriété d'être invariant par translation, tout comme le périodogramme classique pour les données régulières et τ est choisi tel que $\sum_{i=1}^n \cos(\omega(t_i - \tau)) \sin(\omega(t_i - \tau)) = 0$. En 1981, Scargle [26] étudie le comportement statistique du périodogramme modifiée de Lomb. Cette méthode porte ainsi le nom de la méthode de Lomb-Scargle. En particulier, il montre que la méthode est exactement équivalente à une méthode des moindres carrés avec fonction de régression $g(\omega, A, B; t) = A \cos(\omega t) + B \sin(\omega t)$. De plus, il est possible d'effectuer un test sur la signficance de la plus haute valeur du périodogramme modifié. Cependant, l'hypothèse sous-jacente est que le nombre de fréquences de Fourier indépendantes est supposé être de $\lceil \frac{n}{2} \rceil$, si n est paire, $\lceil \frac{n-1}{2} \rceil$, si n est impair, comme dans le cas d'observations régulières. Ceci n'est pas vérifié pour des observations irrégulières. Plus tardivement, Zechmeister et Kürster [35], en 2008, proposent le périodogramme de Lomb-Scargle généralisé, en considérant la fonction de régression $s(t) = A \cos(\omega t) + B \sin(\omega t) + \mu$. Cette méthode est donc équivalente à l'analyse harmonique, qui sera étudiée au chapitre 2. Ferraz-Mello [8] propose, en 1981, une analyse harmonique, où les fonctions de régressions sont orthogonalisées par le procédé de Gram-Schmidt. Il utilise aussi une méthode de soustraction d'une composante sinusoidale détectée dans les données, de façon à pouvoir identifier d'éventuelles autres fréquences présentes dans les données. En 1983, Marquardt et Acuff [19] s'intéresse à la fonction d'autocorrélation, et définissent la méthode "Direct Quadratic Spectrum Estimation" (DQSE). C'est une estimation basée sur une pondération des coefficients d'autocorrélation

$$\frac{1}{T} \sum_{i < j = 1}^n w_{ij} (y_i - \bar{y})(y_j - \bar{y}),$$

où y_i , $i = 1, \dots, n$ sont les observations, \bar{y} la moyenne des observations et T le temps total d'observation qui peut varier en fonction du modèle d'échantillonnage irrégulier. Les poids w_{ij} dépendent d'un

facteur qui est en fonction de la distance entre les décalages δ_{lk} et le décalage δ_{ij} considéré et une fonction cosinus avec une certaine fréquence. Ces poids sont donc adaptés à l'échantillonnage particulier. En 1986, Roberts et al. [25] adapte l'algorithme "clean", utilisé dans la reconstruction d'images, au cas de l'analyse spectrale. Il utilise cette méthode pour retirer les artefacts apparaissant dans le périodogramme. En 1995, Grant Foster [10] introduit la méthode du "cleanest algorithm". Il considère le périodogramme classique et lui retire les faux pics dans le périodogramme apparaissant à cause de l'échantillonnage irrégulier. En ce qui concerne l'étude de la distribution asymptotique de l'estimateur de la fréquence pour des observations régulières avec des erreurs suivant un processus Gaussien indépendant, Walker [32] l'a traitée en 1971. En 1973, Walker [33] donne la distribution asymptotique de la fréquence pour le cas d'un signal sinusoïdal observé de manière régulière avec des erreurs corrélées. Il considère un processus autorégressif AR infini pour les erreurs. En 1981, Kovacs [15] propose une approximation de la variance asymptotique pour l'estimateur de la fréquence, lorsque les observations régulières, qui est la même que celle trouvée par Walker [32]. En 2000, Hall et al. [13] proposent une méthode non paramétrique pour une estimation simultanée de la forme de la fonction ainsi que de la période dans le cas d'observations irrégulières avec des erreurs indépendantes. Les auteurs donnent la distribution asymptotique de l'estimateur \hat{P}_n de la période P_0 ,

$$n^{\frac{3}{2}} (\hat{P}_n - P_0) \sim \mathcal{N}(0, \tau^2),$$

avec

$$\tau^2 = 12\sigma^2 P_0^3 \eta^2 \left\{ \int_0^{P_0} g'(u)^2 du \right\}^{-1}, \text{ et } \eta = \lim_{i \rightarrow \infty} \mathbb{E}[T_i - T_{i-1}],$$

où T_i est le i -ème temps aléatoire d'observation. Au chapitre 2, nous verrons une explication détaillée de cette méthode. Au chapitre 3, les méthodes de l'analyse harmonique, non paramétrique, du périodogramme en amplitude (méthode de Deeming) et de l'autocorrélation sont appliquées aux données du catalogue A de la mission Hipparcos. Cette étape permettra, entre autres, de sélectionner les meilleures méthodes. Au chapitre 4 nous aborderons une généralisation de cette distribution lors d'erreurs suivant un processus Gaussien corrélé, ainsi que dans le cas où la fonction de régression est sinusoïdale. Nous nous intéresserons aussi à l'étude de la variance asymptotique lorsque le signal g comporte des discontinuités. L'intérêt pour ce genre de signal provient de l'article [16], où les auteurs opèrent une régression sur des observations de transit d'exoplanètes et où la fonction de régression à ajuster est un signal de type boîte. Finalement le chapitre 5 sera dédié à la comparaison des méthodes de l'analyse harmonique, non paramétrique et du périodogramme en amplitude sur un ensemble fini d'observations, et cela pour différents signaux périodiques ainsi que différents modèles d'échantillonnage pour les temps d'observations.

Chapitre 2

Méthodes de recherche de fréquences cachées

Dans la section 1.5 sur la théorie des séries temporelles, nous avons vu que les méthodes dites classiques de recherche de fréquences ne sont pas plus efficaces lorsque l'échantillonnage est irrégulier. L'estimateur de la fréquence a tendance à prendre des valeurs qui sont des alias de la fréquence. Dans ce chapitre, nous présentons les méthodes de Deeming (D), de l'analyse harmonique (AH), la régression non paramétrique (NP), ainsi qu'une méthode d'estimation combinant l'analyse harmonique et l'estimation de la forme du signal. Cette dernière est définie dans ce chapitre, mais elle ne sera pas considérée dans les chapitres suivants. En effet, si la fréquence est bien estimée par la méthode AH, et si l'échantillonnage est assez dense, on obtient facilement une bonne estimation de la forme de la fonction. Ainsi les considérations seront effectuées sur l'estimateur de la fréquence et non sur l'estimateur de la fonction. La performance asymptotique de ces méthodes sera donnée dans le chapitre 4 sur la théorie asymptotique. Considérons à nouveau le modèle des observations en temps discret (1.3)

$$Y_i = g_i + \varepsilon_i, i = 1, \dots, n, \quad (2.1)$$

où $g_i = g(t_i)$ est la fonction périodique g , de période P_0 , observée aux temps t_i . Le processus stochastique en temps discret $\varepsilon_i \sim \mathcal{N}(0, \sigma^2), i = 1, \dots, n$ est un processus Gaussien stationnaire et ergodique (section 1.6) avec fonction d'autocovariance $\gamma(|t_i - t_j|) \in \mathbb{R}(1.6), i, j = 1, \dots, n$. Pour toutes ces méthodes, nous considérons un échantillon (y_1, \dots, y_n) de taille n , contenant une fréquence inconnue ω_0 . Pour l'illustration de ces méthodes, considérons le signal périodique sinusoidal

$$g(t) = A \cos(\omega_0 t) + B \sin(\omega_0 t) + \mu, \quad (2.2)$$

de période $P_0 = 10$, c'est-à-dire de fréquence $\omega_0 = \frac{\pi}{5}$, avec une amplitude $R = \sqrt{A^2 + B^2} = 1$ et $A = B = \frac{\sqrt{2}}{2}$ et $\mu = 0$. L'ensemble (g_1, \dots, g_n) est le signal aux temps $(t_1, \dots, t_n), n = 80$, avec un échantillonnage suivant un processus de Poisson $\mathcal{P}(1)$. Le processus stochastique en temps discret $\varepsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), i = 1, \dots, n$ est un processus Gaussien stationnaire et ergodique iid (section 1.6). Les observations irrégulières $y_i = g(t_i) + \varepsilon_i, i = 1, \dots, n$, sont représentées dans la figure 2.1. L'approche classique était d'estimer la densité spectrale par le périodogramme de Schuster ([27] et (1.40)). Pour le calcul du périodogramme, on considère la méthode de Deeming, une estimation du périodogramme en amplitude (1.42).

2.1 L'analyse harmonique

L'analyse harmonique est l'étude de la présence d'une composante sinusoidale $A_g \cos(\omega_0 t) + B_g \sin(\omega_0 t) + \mu_g$ dans un signal $g(t)$ périodique donné de période P_0 . En faisant référence à la théorie des sé-

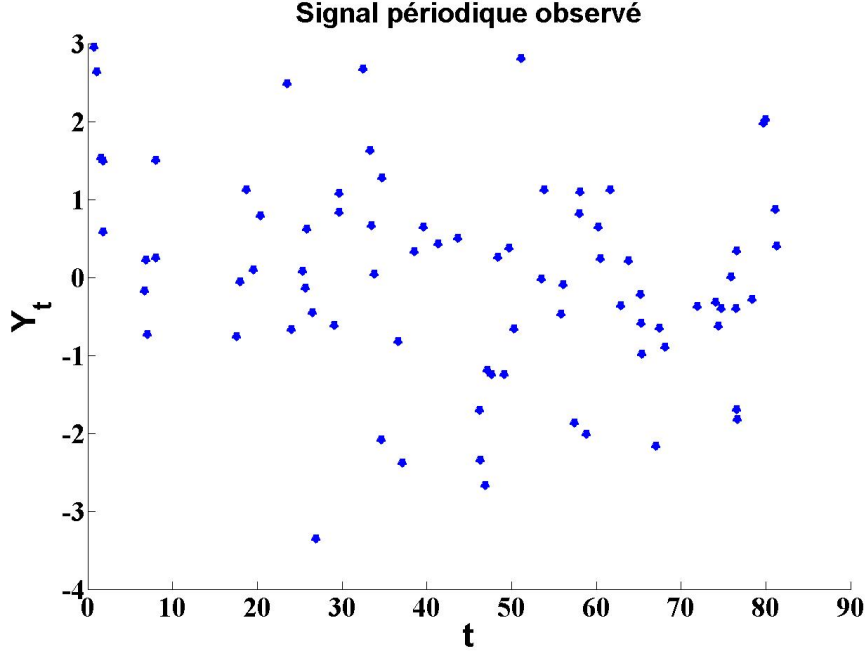


FIG. 2.1 – Observations irrégulières du signal (2.2) de fréquence $\omega_0 = \frac{\pi}{5}$.

ries de Fourier (1.9), posons $c_{g,1} = A_g + iB_g$, et $c_{g,-1} = A_g - iB_g$. On peut donc remarquer que $c_{g,1} = \mathcal{TF}\{g\}(\omega_0)$, $c_{g,-1} = \mathcal{TF}\{g\}(-\omega_0)$, $A_g = \frac{c_{g,1} + c_{g,-1}}{2}$ et $B_g = \frac{c_{g,1} - c_{g,-1}}{2i}$. Définissons

$$s_g(t) = A_g \cos(\omega_0 t) + B_g \sin(\omega_0 t) + \mu_g, \quad (2.3)$$

et

$$r_g(t) = g(t) - s_g(t). \quad (2.4)$$

La théorie de Fourier nous dit que

$$r_g(t) = \sum_{k=-\infty}^{-2} c_k \exp(ik\omega_0 t) + \sum_{k=2}^{\infty} c_k \exp(ik\omega_0 t).$$

Le modèle (2.1) devient

$$Y_i = s_g(t_i) + r_g(t_i) + \varepsilon_i, i = 1, \dots, n. \quad (2.5)$$

Dans l'exemple (1.10), $g(t) = s_g(t)$ et $r_g(t) = 0$, avec $A_g = A = \frac{\sqrt{2}}{2}$, $B_g = B = \frac{\sqrt{2}}{2}$ et $\mu_g = \mu = 0$. Lorsqu'il s'agit d'estimer $(\omega_0, A_g, B_g, \mu_g)$ par moindres carrés, on introduit le critère

$$SC(\omega, A, B, \mu) = \sum_{i=1}^n (Y_i - s(t_i))^2, \quad (2.6)$$

avec la fonction de régression

$$s(t) = A \cos(\omega t) + B \sin(\omega t) + \mu, \quad (2.7)$$

et ceci pour chaque $\omega \in \mathbb{R}$. En pratique, nous limitons les valeurs de la fréquence ω possibles à un intervalle compact et considérons (2.6) pour une fréquence ω donnée. L'ajustement de la fonction de régression (2.7) fournit l'estimateur

$$\widehat{s}(t) = \widehat{A}_n \cos(\omega t) + \widehat{B}_n \sin(\omega t) + \widehat{\mu}_n, \quad (2.8)$$

de la fonction $s(t)$, où \widehat{A}_n , \widehat{B}_n et $\widehat{\mu}_n$ sont les solutions du système d'équation (2.9)

$$\begin{cases} \sum_{i=1}^n Y_i \cos(\omega t_i) - \widehat{A}_n \sum_{i=1}^n \cos(\omega t_i)^2 - \widehat{B}_n \sum_{i=1}^n \cos(\omega t_i) \sin(\omega t_i) - \widehat{\mu}_n \sum_{i=1}^n \cos(\omega t_i) = 0, \\ \sum_{i=1}^n Y_i \sin(\omega t_i) - \widehat{A}_n \sum_{i=1}^n \cos(\omega t_i) \sin(\omega t_i) - \widehat{B}_n \sum_{i=1}^n \sin(\omega t_i)^2 - \widehat{\mu}_n \sum_{i=1}^n \sin(\omega t_i) = 0, \\ \sum_{i=1}^n Y_i - \widehat{A}_n \sum_{i=1}^n \cos(\omega t_i) - \widehat{B}_n \sum_{i=1}^n \sin(\omega t_i) - n\widehat{\mu}_n = 0. \end{cases} \quad (2.9)$$

En résolvant le système (2.9) de 3 équations à 3 inconnues, et en posant que

$$\begin{aligned} CC &= n \sum_{i=1}^n \cos(\omega t_i)^2 - \left(\sum_{i=1}^n \cos(\omega t_i) \right)^2, \\ SS &= n \sum_{i=1}^n \sin(\omega t_i)^2 - \left(\sum_{i=1}^n \sin(\omega t_i) \right)^2, \\ CS &= n \sum_{i=1}^n \cos(\omega t_i) \sin(\omega t_i) - \left(\sum_{i=1}^n \cos(\omega t_i) \right) \left(\sum_{i=1}^n \sin(\omega t_i) \right), \\ CY &= n \sum_{i=1}^n Y_i \cos(\omega t_i) - \left(\sum_{i=1}^n Y_i \right) \left(\sum_{i=1}^n \cos(\omega t_i) \right), \\ SY &= n \sum_{i=1}^n Y_i \sin(\omega t_i) - \left(\sum_{i=1}^n Y_i \right) \left(\sum_{i=1}^n \sin(\omega t_i) \right). \end{aligned}$$

on obtient que

$$\widehat{A}_n(\omega) = \widehat{A}_n = \frac{CY}{CC} - \frac{CS}{CS^2 - CCSS} \left[-SY + \frac{CS}{CC} CY \right], \quad (2.10)$$

$$\widehat{B}_n(\omega) = \widehat{B}_n = \frac{SYCC}{CCSS - CS} - \frac{CSCC}{CS^2 - CCSS} CY, \quad (2.11)$$

$$\widehat{\mu}_n(\omega) = \widehat{\mu}_n = \bar{Y} - \widehat{A}_n \frac{\sum_{i=1}^n \cos(\omega t_i)}{n} - \widehat{B}_n \frac{\sum_{i=1}^n \sin(\omega t_i)}{n}. \quad (2.12)$$

Dans le cas où les observations sont régulières $t_i = i$, $i = 1, \dots, n$, et en considérant les fréquences de Fourier $\omega_k = \frac{2\pi k}{n}$, $k = 0, \dots, n-1$, on a que $CS = 0$, $CC = \sum_{i=1}^n \cos(\omega_k t_i)^2$, $SS = \sum_{i=1}^n \sin(\omega_k t_i)^2$, $CY = \sum_{i=1}^n Y_i \cos(\omega_k t_i)$ et $SY = \sum_{i=1}^n Y_i \sin(\omega_k t_i)$. Ainsi (2.12) devient

$$\widehat{A}_n = \frac{\sum_{i=1}^n Y_i \cos(\omega_k t_i)}{\sum_{i=1}^n \cos(\omega_k t_i)^2}, \quad (2.13)$$

$$\widehat{B}_n = \frac{\sum_{i=1}^n Y_i \sin(\omega_k t_i)}{\sum_{i=1}^n \sin(\omega_k t_i)^2}, \quad (2.14)$$

$$\widehat{\mu}_n = \bar{Y} - \widehat{A}_n \frac{\sum_{i=1}^n \cos(\omega_k t_i)}{n} - \widehat{B}_n \frac{\sum_{i=1}^n \sin(\omega_k t_i)}{n}. \quad (2.15)$$

Définissons l'estimateur minimisant (2.6), parmi toutes les fréquences ω considérées, par

$$\hat{\omega}_n^{\text{AH}} = \operatorname{argmin}_{\omega} \text{SC}(\omega, \hat{A}_n, \hat{B}_n, \hat{\mu}_n). \quad (2.16)$$

Le théorème 4.17 sur la consistance de l'estimateur de la fréquence $\hat{\omega}_n^{\text{AH}}$ montre que cette méthode appliquée à un signal périodique g quelconque de période P_0 , tel que $g|_{[0, P_0)} \in L^2([0, P_0))$ est valable. La somme des carrés des résidus est donc évaluée pour chaque valeur de ω sur un intervalle raisonnable. Pour chaque fréquence ω , les estimateurs \hat{A}_n , \hat{B}_n et $\hat{\mu}_n$ sont évalués, ce qui permet de définir le meilleur ajustement possible au sens des moindres carrés pour chaque valeur de ω . Ainsi la fréquence minimisant $\text{SC}(\omega)$ sur cet intervalle est l'estimateur $\hat{\omega}_n^{\text{AH}}$ de la vraie fréquence inconnue ω_0 . Pour décider quelle fréquence ω est effectivement un bon estimateur de ω_0 , nous considérons un test F

$$T = \frac{\left(\text{SC}(\omega_0, A_g = 0, B_g = 0, \hat{\mu}_n) - \text{SC}(\hat{\omega}_n^{\text{AH}}, \hat{A}_n, \hat{B}_n, \hat{\mu}_n) \right) / dl_1}{\text{SC}(\hat{\omega}_n^{\text{AH}}, \hat{A}_n, \hat{B}_n, \hat{\mu}_n) / dl_2}. \quad (2.17)$$

Asymptotiquement, $T \sim F_{dl_1, dl_2}$, avec $dl_1 = 3$ et $dl_2 = n - 4$. L'hypothèse nulle $H_0 : A_g = B_g = 0$ est rejetée si $T > qF_{3, n-4}(95\%)$. On applique la méthode aux données de la figure 1.4. La figure 2.2 représente la somme des carrés des résidus (4.77). La figure 2.3 représente le test F (2.17). Le minimum de la figure 2.2 donne la même conclusion que le maximum dans la figure 2.3. En effet, les auteurs de [35] montrent que si l'on normalise (2.6), alors cette méthode est équivalente au test F (2.17) à une constante près. Cette constante est simplement le rapport des degrés de liberté, ce qui ne change rien quant au choix du maximum du test F. Ces degrés de liberté sont importants dans la comparaison de ce test au quantile $(1 - \alpha)$ de la loi F de Fisher. Si l'hypothèse nulle est rejetée, alors le maximum observé correspond vraisemblablement à une fréquence présente dans le signal observé.

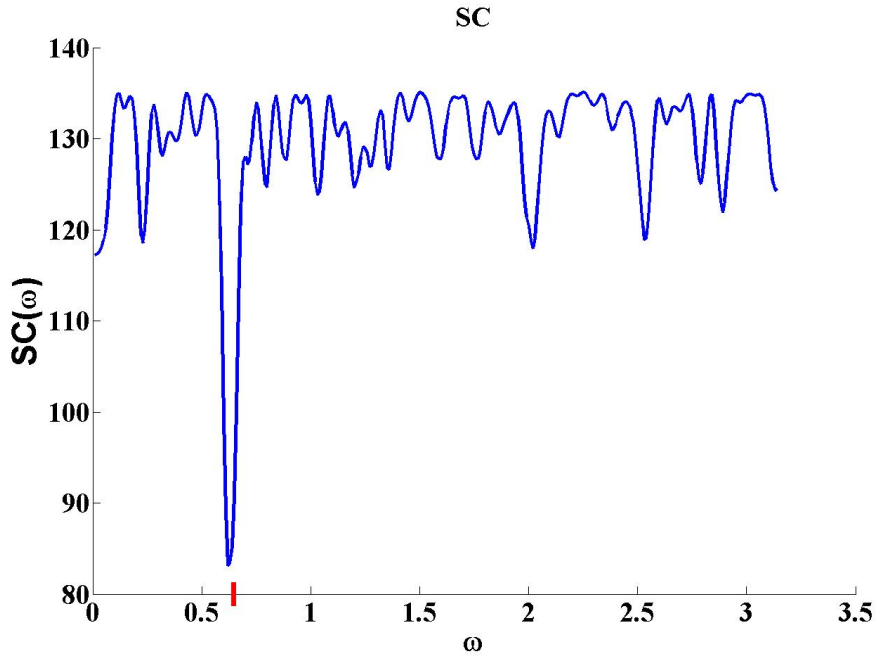


FIG. 2.2 – Somme des carrés (4.77) pour les observations irrégulières, figure 2.1, du signal (2.2) de fréquence $\omega_0 = \frac{\pi}{5} \simeq 0.628$ (représentée par une barre verticale rouge sur l'axe des abscisses). L'estimateur $\hat{\omega}_n^{\text{AH}}$ vaut 0.62

2.2 Méthode de Deeming

La méthode de Deeming est une estimation du périodogramme en amplitude. Tout comme l'analyse harmonique, il s'agit de l'étude de la présence d'une composante sinusoidale de fréquence ω_0 dans un

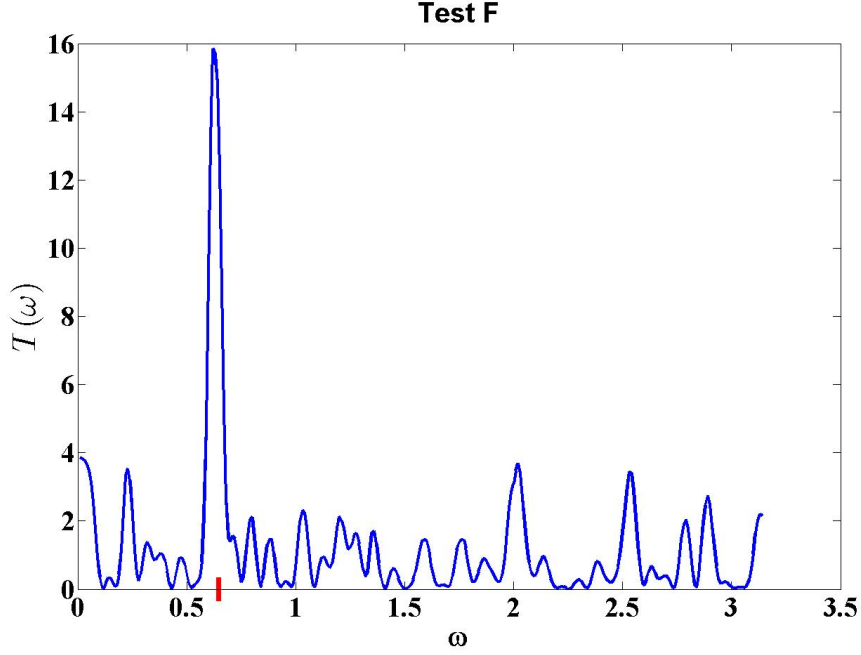


FIG. 2.3 – Test F (2.17) pour les observations irrégulières, figure 2.1, du signal (2.2) de fréquence $\omega_0 = \frac{\pi}{5} \simeq 0.628$ (représentée par une barre verticale rouge sur l’axe des abscisses). L’estimateur $\hat{\omega}_n^{\text{AH}}$ vaut 0.62, et le test F rejette H_0 .

signal périodique. Avec les mêmes hypothèses, considérées dans la section précédente 2.1, sur le modèle des observations (2.1), l’estimateur $\hat{\omega}_n^{\text{D}}$ de la fréquence ω_0 est tel que

$$\hat{\omega}_n^{\text{D}} = \operatorname{argmax}_{\omega} P(\omega), \quad (2.18)$$

où

$$P(\omega) = \frac{2}{n} \sqrt{\left(\sum_{i=1}^n Y_i \cos(\omega t_i) \right)^2 + \left(\sum_{i=1}^n Y_i \sin(\omega t_i) \right)^2} \quad (2.19)$$

est l’estimation du périodogramme en amplitude, qui a déjà été introduit dans la section 1.11.

2.3 Remarques sur les méthodes de l’analyse harmonique et de Deeming

Comme mentionné au début des sections 2.1 et 2.2, l’analyse harmonique et la méthode de Deeming procède à une étude de la présence d’une composante sinusoidale de fréquence ω_0 . A présent, considérons une fonction g périodique de période P_0 , tel que $g \in L^1([0, P_0])$. Posons $h(t) = g(t) - s(t) = r(t)$, où $s(t)$ correspond à une harmonique de la fréquence ω_0 . La fonction $h(t)$ est toujours périodique de période P_0 , mais ne contient plus la composante sinusoidale de fréquence ω_0 . Au chapitre 4 sur la théorie asymptotique, le lemme 2 de la démonstration du théorème 9 montre, en particulier, que $s(t)$ et $r(t)$ sont orthogonaux au sens des moindres carrés. Ainsi, effectuer une analyse harmonique sur le signal $h(t)$ ne permettra pas de recouvrir la fréquence du signal ω_0 , mais seulement un de ses multiples, en l’occurrence celui dont le coefficient de Fourier est le plus prédominant. Il y a donc une infinité de fonctions périodiques de type $h(t)$ qui ont période P_0 , mais où l’analyse harmonique et la méthode de Deeming ne peuvent pas détecter la période.

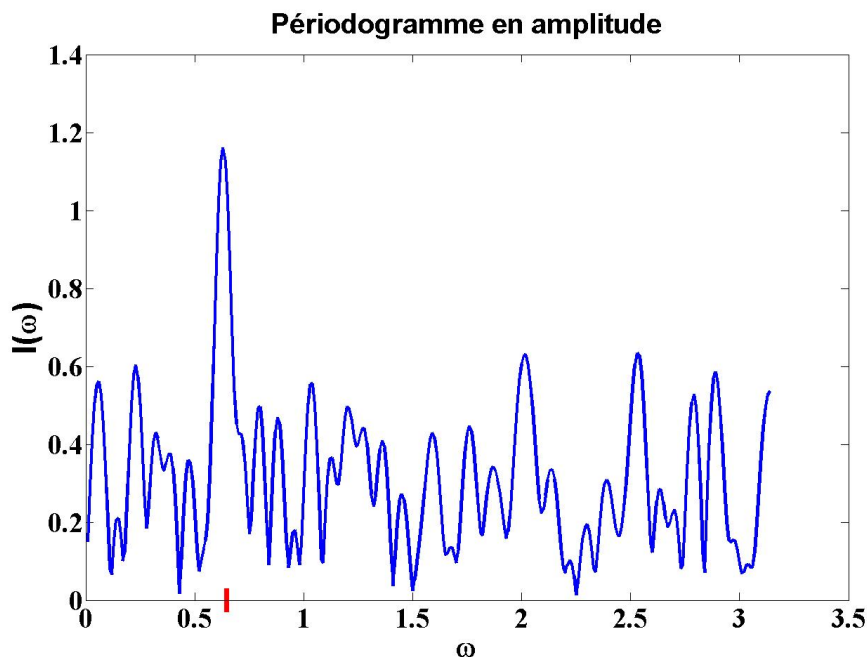


FIG. 2.4 – Périodogramme en amplitude pour les observations irrégulières, figure 2.1, du signal (2.2) de fréquence $\omega_0 = \frac{\pi}{5} \simeq 0.628$ (représentée par une barre verticale rouge sur l’axe des abscisses). La fréquence estimée vaut 0.63.

2.4 Estimation par régression non paramétrique

Le modèle des observations est (2.1). Lorsqu’il s’agit d’estimer ω_0 par moindres carrés, on introduit le critère

$$SC(\omega, g) = \sum_{i=1}^n (Y_i - g(t_i))^2, \quad (2.20)$$

où g est un paramètre fonctionnel périodique de période $P = \frac{2\pi}{\omega}$. Ce paramètre g est estimé en effectuant un repliement en phase des observations, puis en effectuant un lissage de ces observations repliées. Cela fournit une estimation $\hat{g}_n(t)$ de la fonction g . A nouveau, l’estimateur $\hat{g}_n(t)$ est déterminé pour une fréquence ω fixe. L’estimateur $\hat{\omega}_n^{\text{NP}}$ est ainsi défini par

$$\hat{\omega}_n^{\text{NP}} = \operatorname{argmin}_{\omega} SC(\omega, \hat{g}_n). \quad (2.21)$$

Lors de l’utilisation de cette idée, qui est due à McDonald, [22], nous avons utilisé un lissage localement linéaire. La méthode du repliement en phase aux données (2.1) produit la figure 2.5 pour la fréquence estimée par la méthode. La figure 2.2 représente la somme des carrés des résidus (2.20). Comme dans la section précédente, on considère une grille raisonnable des valeurs de la période à tester. A ce moment là, une méthode de lissage peut être appliquée aux données translatées dans l’intervalle $[0, P)$. En utilisant une méthode de lissage localement linéaire, on obtient la figure 2.6 avec les données de la figure 2.5. Cette méthode souffre du problème de la multiplicité de la période. En effet, la somme des carrés des résidus (2.20) évaluée en $\omega = \frac{\pi}{P}$ donnera aussi un minimum local. De même que $\omega = \frac{2\pi}{3P}, \dots, \omega = \frac{2\pi}{kP}$, avec $k < \frac{T}{P}$ et T le temps total d’observation du signal. Suivant la valeur du rapport signal bruit, un multiple de la période pourrait être mieux détecté que la période elle-même. Cette propriété provient de l’estimation de la forme du signal par les données. Pour une valeur de la période égale à $2P$, le signal estimé sera la répétition de deux fois le signal estimé sur une période. Dans l’exemple que nous

considérons, la somme des carrés (2.20) atteint une valeur minimale pour la fréquence $\hat{\omega}_n^{\text{NP}} = 0.63$, ce qui correspond à une période de presque 10. Cette propriété de la multiplicité de la période apparaît dans la figure 2.7, où les valeurs de la somme des carrés des résidus (2.20), correspondant à une fréquence proche d'une fraction $\frac{1}{k}\omega_0$ de la fréquence ω_0 , sont représentées par des étoiles en rouge, pour $k = 1, \dots, 5$. Dans cet exemple $\frac{1}{4}\omega_0$ est proche du minimum global. Dans [13], les auteurs relèvent ce problème. Pour l'analyse de la variance asymptotique, ils supposent que la minimisation de (2.20) est effectuée dans un intervalle autour de la période P_0 . Ce problème n'apparaît pas avec l'analyse harmonique. Pour une valeur de la fréquence $\frac{\omega}{2}$ (équivalent à une période de valeur $2P$), la fonction de régression (2.7) aura une période de $2P$ et ne sera pas la répétition de deux fois la fonction de période P .

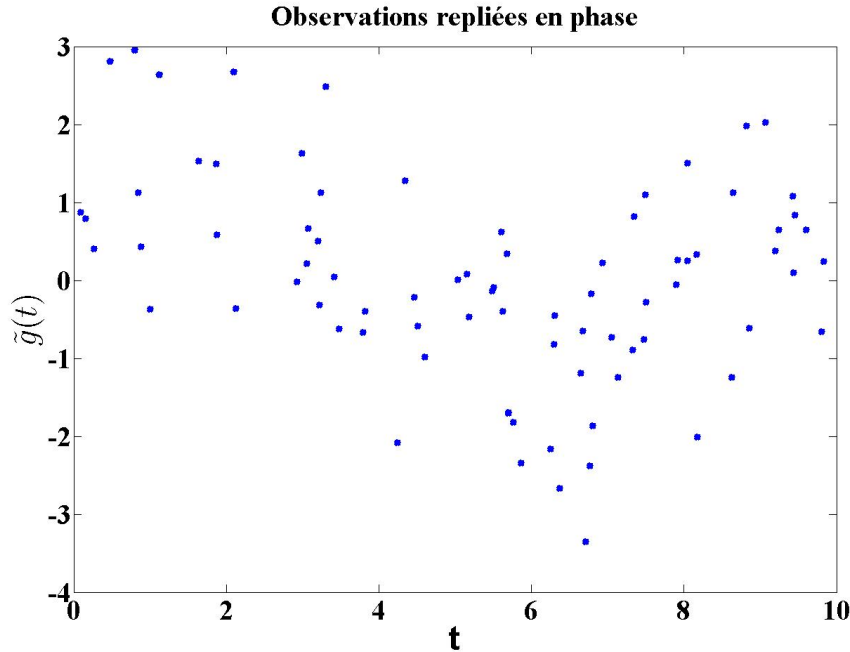


FIG. 2.5 – Repliement en phase des données, avec une période estimée $\hat{P}_n^{\text{NP}} = 10$, pour les observations irrégulières, figure 2.1, du signal (2.2) de période $P_0 = 10$.

2.5 Mariage entre l'analyse harmonique et la régression non paramétrique

Contrairement à la méthode non paramétrique, cette méthode est structurée en deux étapes. La première étape consiste à estimer la fréquence du signal par la méthode de l'analyse harmonique. La deuxième étape est l'estimation de la forme du signal, par un repliement en phase des données avec la période estimée par l'analyse harmonique (voir la méthode repliement en phase de la section précédente). Cette méthode permet d'obtenir une bonne estimation de la fréquence (voir les théorèmes 8 et 9 concernant la consistance de l'estimateur de la fréquence et la variance asymptotique de la fréquence respectivement) sans le problème de la multipériodicité comme dans la méthode non paramétrique. Avec l'exemple du signal de période $P_0 = 10$, la méthode de l'analyse harmonique fournit une fréquence estimée $\hat{\omega}_n^{\text{AH}} = 0.6104$. La figure 2.8 montre la fonction estimée $\hat{g}_n(t)$ sur $[0, P_0)$, à l'aide de la fréquence estimée $\hat{\omega}_n^{\text{AH}}$. Cette méthode est simplement proposée comme étant une méthode efficace pour l'estimation de la forme de la fonction g . Mais elle ne sera pas utilisée, car nous nous intéressons à l'estimation de la fréquence ω_0 et non à la forme de la fonction. Le tableau 2.1 présente une synthèse des méthodes de l'analyse harmonique et de l'estimation par régression non paramétrique.

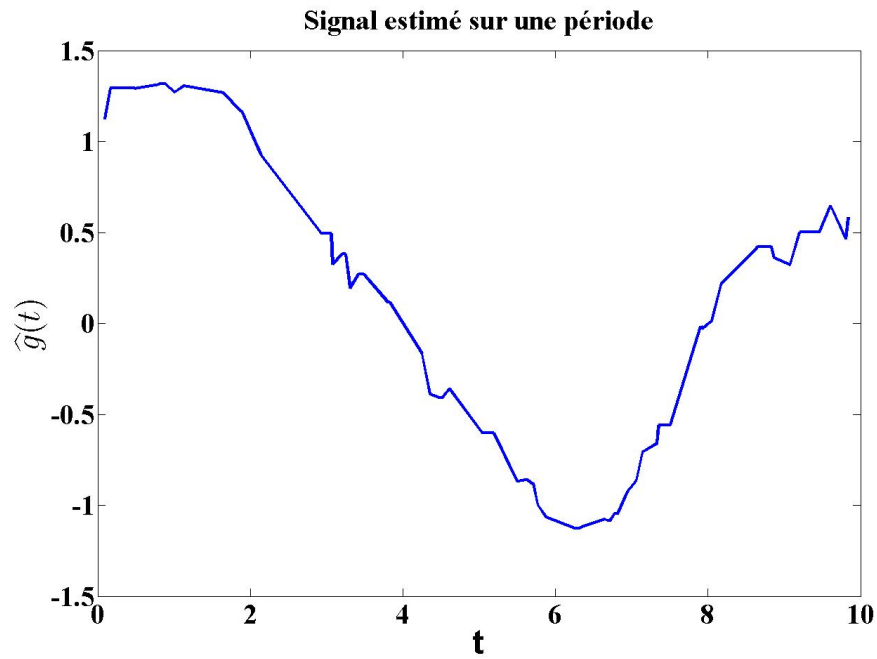


FIG. 2.6 – Lissage localement linéaire des données pour les observations irrégulières, figure 2.1 du signal (2.2) de période $P_0 = 10$.

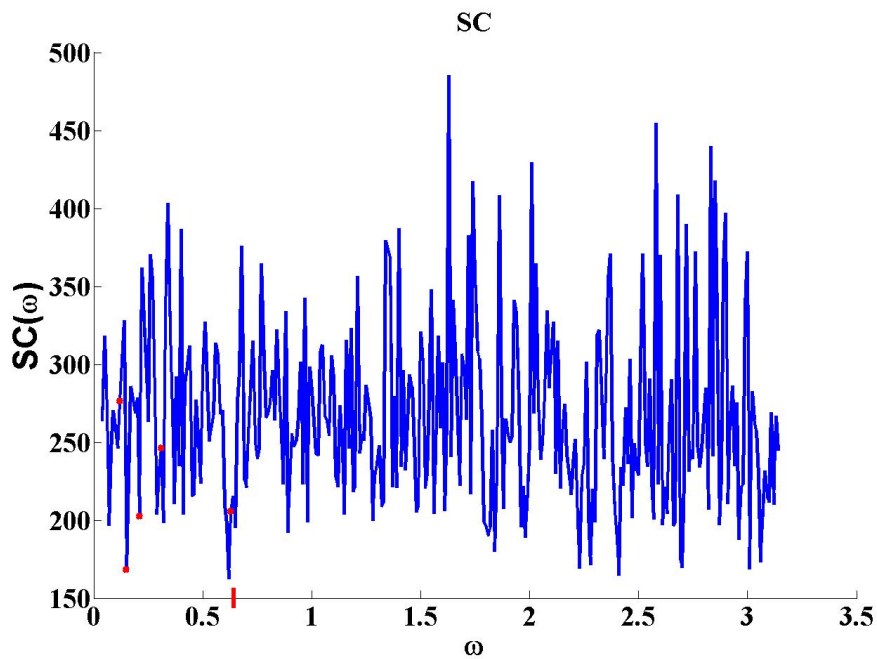


FIG. 2.7 – Somme des carrés (2.20) pour les observations irrégulières, figure 2.1, du signal (2.2) de période $P_0 = 10$. Le symbole étoile en rouge indique les valeurs de la somme des carrés des résidus (2.20) correspondant à une fréquence proche d’une fraction $\frac{1}{k}\omega_0$ de la fréquence ω_0 .

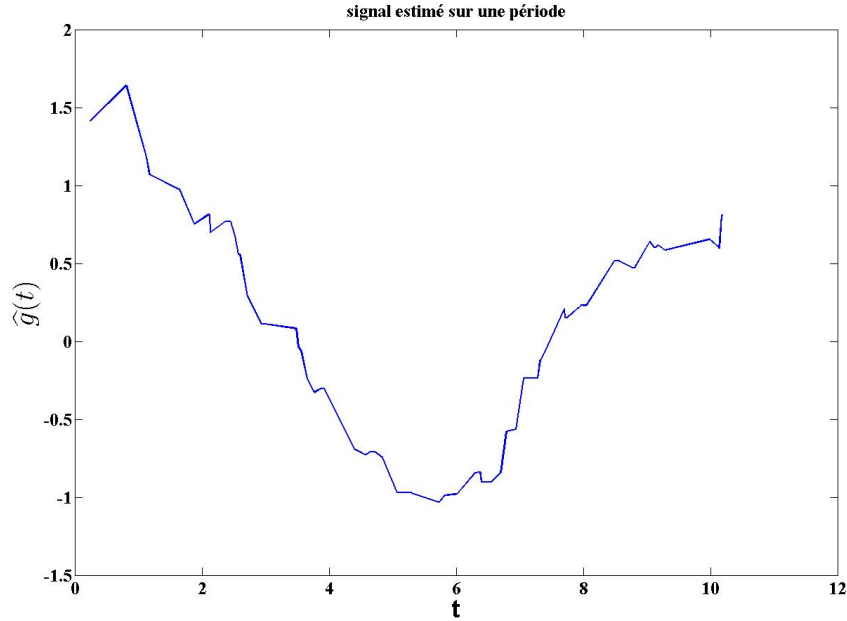


FIG. 2.8 – Estimation $\hat{g}_n(t)$ de la fonction $g(t)$ pour les observations irrégulières, figure 2.1, du signal 2.2 de période $P_0 = 10$, à l'aide de la fréquence estimée $\hat{\omega}_n^{\text{AH}}$.

2.6 Récapitulatif des méthodes de régressions

Dans ce chapitre, nous avons présenté l'analyse harmonique (AH), la méthode de Deeming (D), la méthode de la régression non paramétrique ainsi qu'une méthode combinant l'analyse harmonique et la régression non paramétrique. Le tableau 2.1 donne un résumé des deux méthodes de régressions, AH et NP.

	Analyse harmonique (AH)	Régression non paramétrique (NP)
Modèle	$Y(t_i) = g_\theta(t_i) + r_\theta(t_i) + \varepsilon(t_i)$	$Y(t_i) = g(t_i) + \varepsilon(t_i)$
Paramètres	$(\omega_0, A_g, B_g, \mu_g)$	(ω_0, g)
Fonction de régression	$s(t) = A \cos(\omega t) + B \sin(\omega t) + \mu$	$g(t)$
Estimateurs	$(\hat{\omega}_n^{\text{AH}}, \hat{A}_n, \hat{B}_n, \hat{\mu}_n)$	$(\hat{\omega}_n^{\text{NP}}, \hat{g}_n(t))$

TAB. 2.1 – Résumé des méthodes de l'analyse harmonique (AH) et de la régression non paramétrique (NP) avec un échantillon de données de taille n . La fonction de régression est définie pour une fréquence ω (ou période P) donnée. Pour la méthode AH, la fonction de régression est ajustée aux données par la minimisation de (2.6). Pour la méthode NP, la fonction de régression est estimée par un lissage des observations repliées en phase.

Chapitre 3

Application des méthodes sur la partie A du catalogue des étoiles variables de la mission Hipparcos

Dans ce chapitre, les méthodes de Deeming (D), de l'analyse harmonique (AH), de la méthode non paramétrique (NP), ainsi que celle basée sur l'autocorrélation (AC) sont appliquées à la partie A du catalogue des étoiles variables de la mission Hipparcos. Le catalogue des étoiles variables de la mission Hipparcos est composé de trois parties :

- La partie A contient les étoiles variables dont la période a été déterminée,
- la partie B contient les étoiles variables avec de grandes amplitudes, les étoiles pouvant être périodiques ou non périodiques, dont les observations ont été prises depuis la Terre par l'AAVSO (American Association of Variable Star Observers). Ces observations ont été effectuées avant et pendant la mission Hipparcos,
- la partie C contient les étoiles variables dont la période, si l'étoile est effectivement périodique, n'a pas pu être déterminée.

Cette application donnera une comparaison de la performance de ces méthodes sur ce type de données. On verra aussi que sur cet ensemble de données, l'hypothèse que les erreurs dans les observations sont indépendantes ne pèjore pas la performance. Pour cela, on utilise la méthode de l'analyse harmonique pondérée qui sera brièvement explicitée dans la partie 3.2.1. Cette étape a permis de sélectionner les méthodes les plus performantes, dans le but d'étudier de manière plus approfondie la distribution de l'estimateur de la fréquence selon différents modèles d'échantillonnage ainsi qu'une étude de la distribution asymptotique de ces estimateurs. Dans ce chapitre, les performances des méthodes seront présentées, en tenant compte de leur période (tableau 3.2). Nous avons aussi considéré les performances de ces quatre méthodes en sélectionnant les étoiles selon leurs classes. Nous avons considéré 10 classes (tableau 3.6).

3.1 Modèle des observations

On suppose que les observations proviennent du modèle

$$Y_i = g(t_i) + \varepsilon_i, \quad i = 1, \dots, n, \quad (3.1)$$

où $g_i = g(t_i)$ est la vraie valeur du signal g au temps t_i et $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ une variable aléatoire normale d'espérance nulle et de variance σ^2 . Nous supposons que les erreurs de mesures commises sont indépendantes les unes des autres. Ainsi, les erreurs ε_i sont considérées i.i.d., pour tout $i = 1, \dots, n$.

3.1.1 Modèle avec des erreurs corrélées

L'hypothèse des erreurs indépendantes deux à deux et indépendantes du signal n'est pas toujours vérifiée. En effet, pour certains types d'étoiles, les erreurs sont proportionnelles à la magnitude. On rappelle (voir section 1.2) que plus la magnitude est élevée, moins l'intensité lumineuse est forte. Illustration dans la figure 3.1. C'est pourquoi, il existe des méthodes prenant en compte les erreurs. L'analyse harmonique pondérée est une de ces méthodes. Cependant, on verra que le gain apporté n'est pas très probant. L'hypothèse de l'indépendance des erreurs ne péjore pas l'efficacité de l'analyse harmonique pour les données de la mission Hipparcos.

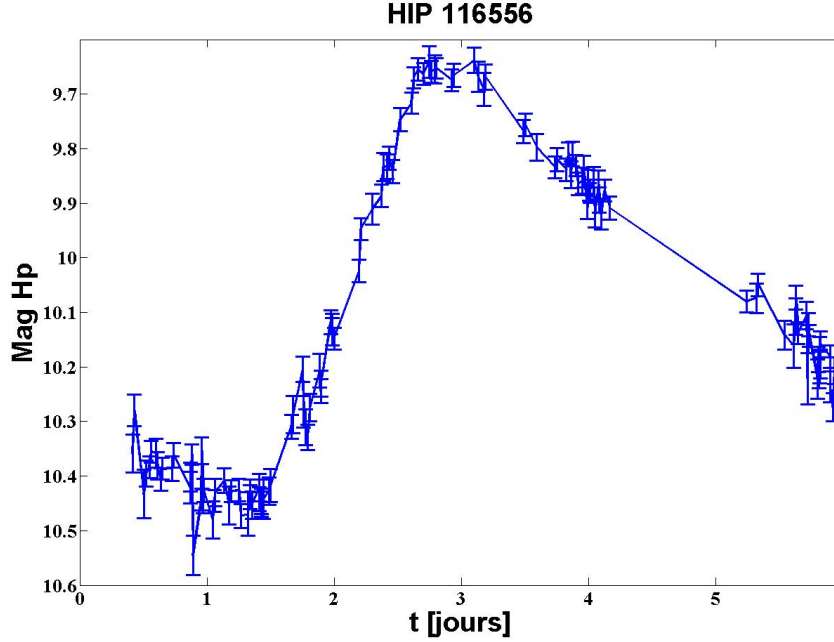


FIG. 3.1 – Erreurs de mesures de l'étoiles HIP 4541 de type RRAB.

3.2 Comparaison des méthodes : analyse harmonique classique et pondérée, autocorrélation, régression non paramétrique et périodogramme en amplitude (Deeming)

Le premier critère CC_1 pour la comparaison de ces méthodes est de considérer une estimation de la fréquence d'un objet comme correcte si elle se trouve dans un voisinage de la fréquence supposée vraie. Nous appliquerons aussi un deuxième critère CC_2 , qui est de considérer aussi un voisinage autour du double de la vraie fréquence. Ici, il s'agit de tenir compte des signaux ayant des périodes internes, comme les binaires à éclipses. Un troisième critère de comparaison CC_3 est de considérer aussi un voisinage autour de la moitié, du tiers et du quart de la vraie fréquence. Ceci pour tenir compte des problèmes de multiplicité de la période rencontrés par les méthodes non paramétrique et d'autocorrélation. Ce voisinage pour la fréquence estimées a été choisi comme étant $2\pi * 0.001$. Ce choix a été fait après avoir tester diverses méthodes (comme par exemple le carré moyen de l'erreur $CME = \mathbb{E} \left((\hat{\omega}_n - \omega_0)^2 \right)$) ainsi que plusieurs voisinages de longueurs différentes.

Les méthodes caractérisent 2683 objets sur les 2712 présents dans la partie A du catalogue des étoiles variables de la mission Hipparcos. Les 29 objets écartés n'ont que trop peu d'observations considérées comme fiables, ce qui ne permet pas d'appliquer une méthode de détection de période.

Les résultats des comparaisons des méthodes sur les données Hipparcos seront représentés par des

diagrammes en colonnes montrant la proportion d'objets que chaque méthode aura correctement caractérisés. La proportion se lit sur l'axe des ordonnées. L'axe des abscisses représente soit l'intervalle dans lequel l'objet a sa période P_0 , (tableau 3.2), soit la classe de l'objet (tableau 3.6).

3.2.1 Comparaison entre l'analyse harmonique et l'analyse harmonique pondérée

L'analyse harmonique pondérée est une généralisation de l'analyse harmonique classique. En effet, on cherche à minimiser la fonction

$$SC_{pond}(\theta) = \frac{1}{2} \sum_{i=1}^n \frac{1}{w_i^2} (Y_i - s_g(t_i))^2, \quad (3.2)$$

où $s_g(t)$ est définie à l'équation (2.3) et les w_i , $i = 1, \dots, n$ sont des poids. Dans l'analyse harmonique classique, $w_i = 1$, $i = 1, \dots, n$. La pondération représente les erreurs associées à chaque mesure de la magnitude. Le tableau 3.1 donne les proportions des 2628 objets correctement caractérisés par les deux méthodes, selon les critères CC_1 et CC_2 . On remarque que, pour les données Hipparcos, l'information sur les erreurs de mesures n'apporte rien de plus à la performance de la méthode.

Critère	AH	AHP
CC_1	0.6149	0.6123
CC_2	0.8512	0.8509

TAB. 3.1 – Proportions de caractérisations corrects des 2683 objets par les méthodes de l'analyse harmonique (AH) et de l'analyse harmonique pondérée (AHP), selon le critère CC_1 et CC_2 .

3.2.2 Comparaison des performances des méthodes de l'analyse harmonique, l'autocorrélation, la régression non paramétrique et la méthode de Deeming

La figure 3.2 montre la comparaison des performances entre les quatre méthodes. On voit clairement la supériorité de la méthode AH, sur l'ensemble des objets. La différence entre le critère de comparaison CC_1 et CC_2 réside dans la forme de la variation de la magnitude apparente des objets de type binaire à éclipses. Ces objets contiennent une période interne, qui est proche de la demi-période. Ainsi, avec le critère CC_2 , on voit que les méthodes D et AH caractérisent une grande partie de ces objets avec leur demi-période. Le critère CC_3 montre la caractéristique de la méthode NP, qui caractérise des objets par un multiple de sa période. Par contre, même avec le critère CC_3 , la méthode basée sur l'autocorrélation ne donne pas de bons résultats sur l'ensemble des objets. Cette comparaison a donc orienté notre choix sur les méthodes de l'analyse harmonique, la méthode non paramétrique et le périodogramme en amplitude. Dans le paragraphe suivant, nous présentons l'analyse de la performance des méthodes sur les données de la partie A du catalogue des étoiles variables de la mission Hipparcos. Notons que l'échantillon d'Hipparcos n'est pas parfait et que nous utilisons donc le terme "fréquence supposée connue" ou "période supposée connue" pour la fréquence ou la période de l'étoile. Les périodes supposées connues de ces objets sont comprises dans l'intervalle (0, 9892) jours.

3.2.3 Analyse de la performance des méthodes sur les données de la partie A du catalogue des étoiles variables de la mission Hipparcos

Nous considérons en premier lieu les objets triés selon leur période. Le critère CC_3 est considéré pour la comparaison des méthodes. Le tableau 3.2 suivant donne le nombre d'objets ayant une période comprise dans l'intervalle correspondant.

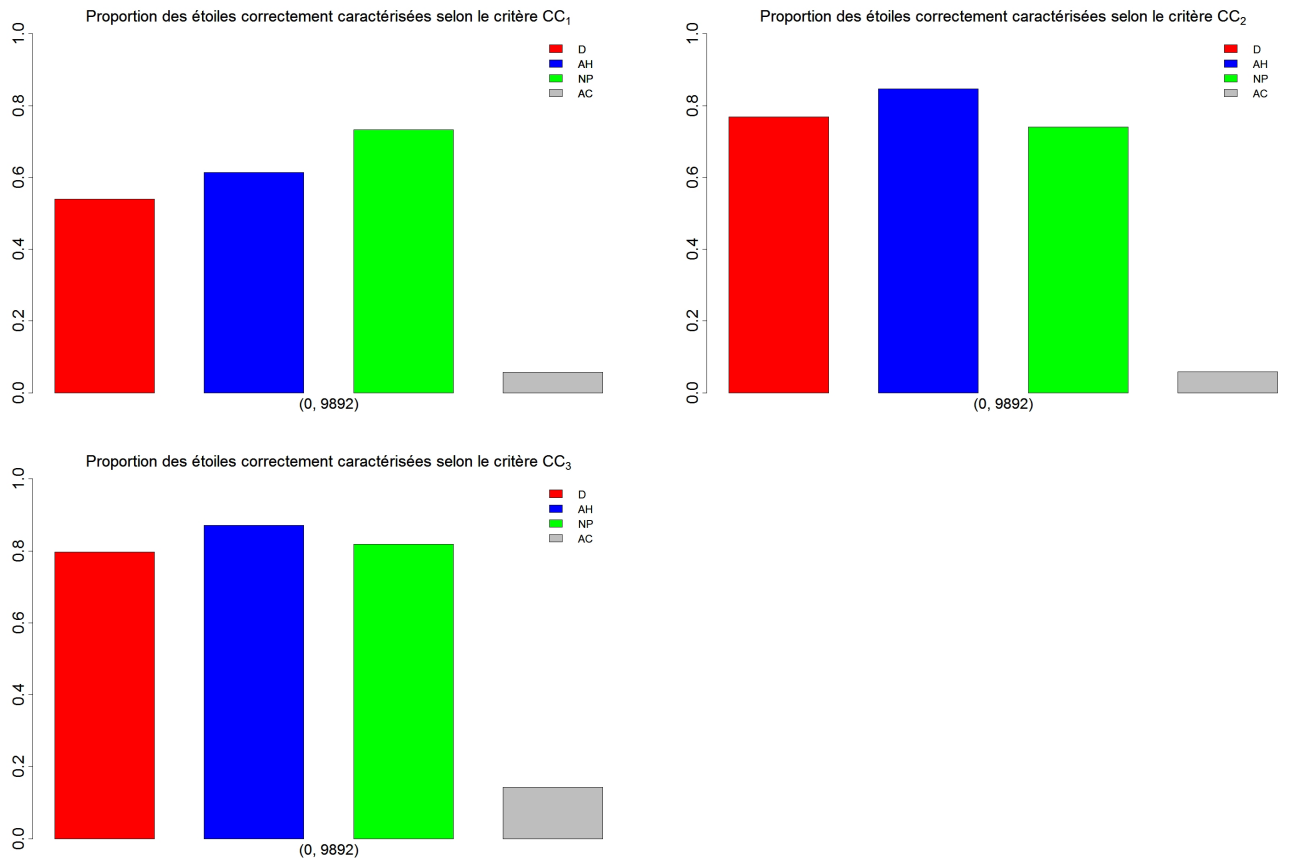


FIG. 3.2 – Graphiques représentant la proportion de caractérisations corrects des 2683 objets, selon le critère CC_1 , CC_2 et CC_3 respectivement.

Intervalles $[b_i, b_s)$	Nombre d'objets avec période $\in [b_i, b_s]$
(0, 1)	783
[1, 10)	1238
[10, 100)	297
[100, 200)	97
[200, 300)	95
[300, 400)	112
[400, 500)	42
[500, 600)	9
[600, 9892)	10

TAB. 3.2 – Objets triés par périodes, b_i est la borne inférieure et b_s la borne supérieure pour la période.

Analyse de la performance des méthodes selon la période de l'objet

Les objets sont triés selon leur période (tableau 3.2). La figure 3.3 confirme que la méthode AH est la plus performante, quelle que soit la période de l'objet, vient ensuite la méthode de Deeming. La méthode NP n'est battue par l'autocorrélation que sur quelques objets ayant des périodes supérieures à 200 jours.

Nous effectuons à présent une analyse sur les objets ayant été correctement caractérisés par une des méthodes, parmi la méthode AH, D et NP. Nous ne considérons plus la méthode basée sur l'autocorrélation étant donné que sa performance est inférieure aux autres. Nous considérons le critère CC_3 pour la caractérisation des objets. Dans les tableaux 3.4 et 3.5, on représente le nombre d'objets caractérisés (C) ou faussement caractérisés (NC) par les méthodes de Deeming (D) et de l'analyse harmonique (AH) (tableau de gauche) et de la méthode l'analyse harmonique et de la méthode non paramétrique (NP) (tableau de droite). Le tableau 3.3 s'interprète comme suit

n objets ayant leur période comprise entre b_i et b_s jours.

		AH		Marges D			AH		Marges NP
		C	NC				C	NC	
D	C	$n_{C,C}$	$n_{C,NC}$	n_C^D	NP	C	$n_{C,C}$	$n_{C,NC}$	n_C^{NP}
	NC	$n_{NC,C}$	$n_{NC,NC}$	n_{NC}^D		NC	$n_{NC,C}$	$n_{NC,NC}$	n_{NC}^{NP}
Marges AH		n_C^{AH}	n_{NC}^{AH}	n	Marges AH		n_C^{AH}	n_{NC}^{AH}	n

TAB. 3.3 – n objets ayant leur période comprise entre b_i et b_s jours, triés selon leur correcte caractérisation (C) ou fausse caractérisation (NC) par les méthodes D et AH (tableau de gauche) et par les méthodes NP et AH (tableau de droite). Le symbole $n_{C,C}$ représente le nombre d'objets correctement caractérisés (C) par les deux méthodes D et AH. Le symbole $n_{C,NC}$ représente le nombre d'objets correctement caractérisés (C) par la méthode D et en même temps faussement caractérisés (NC) par la méthode AH. La dernière colonne du tableau (Marges D) représente le nombre n_C^D d'objets caractérisés par la méthode D (ligne C) et n_{NC}^D le nombre d'objets faussement caractérisés par la méthode D (ligne NC). La dernière ligne du tableau (Marges AH) représente le nombre d'objets correctement caractérisés par la méthode AH (colonne C) et faussement caractérisés par la méthode D (colonne NC). Les symboles $n_{NC,C}$, $n_{NC,NC}$, n_C^{AH} et n_{NC}^{AH} se définissent de manière équivalente pour la méthode AH. Ainsi le nombre total d'objets $n = n_C^D + n_{NC}^D = n_C^{AH} + n_{NC}^{AH} = n_C^{NP} + n_{NC}^{NP}$. Le tableau de droite, concernant la comparaison des performances des méthodes NP et AH, s'interprète de la même manière.

On représente ainsi la répartition par méthode des objets. La comparaison des méthodes D et AH montre que la méthode AH est plus efficace que D. On remarque aussi que les objets caractérisés par l'une ou l'autre méthode ne sont pas forcément les mêmes. Cependant, les deux méthodes caractérisent en commun une majorité des objets. En ce qui concerne la comparaison entre NP et AH, la méthode AH a une meilleure performance que NP.

Analyse de la performance des méthodes selon la classe de l'étoile

Dans cette section, nous considérons la performance de ces méthodes sur quelques types d'étoiles, énumérées dans le tableau 3.6. Les figures 3.4 et 3.5 montrent les courbes lumières approximées par des séries de Fourier, pour chacun de ces types d'étoiles. Ce sont les séries de Fourier finies des observations des objets, dont les observations de la courbe lumière ont été repliées en phase (figures 3.6 et 3.7). La série de Fourier pour le type EA demande un nombre élevé d'harmoniques. De plus le coefficient de Fourier correspondant à la fréquence ω_0 n'est que faiblement présent dans ce type de signal. Tandis que les séries de Fourier des autres types ne demandent qu'entre un et quatre harmoniques, avec une harmonique correspondant à la fréquence fondamentale bien plus fortement présente.

Nous effectuons à présent une analyse sur les objets ayant été correctement caractérisés par une des méthodes, AH, D ou NP. Nous considérons à nouveau le critère CC_3 pour la caractérisation des objets selon leur type.

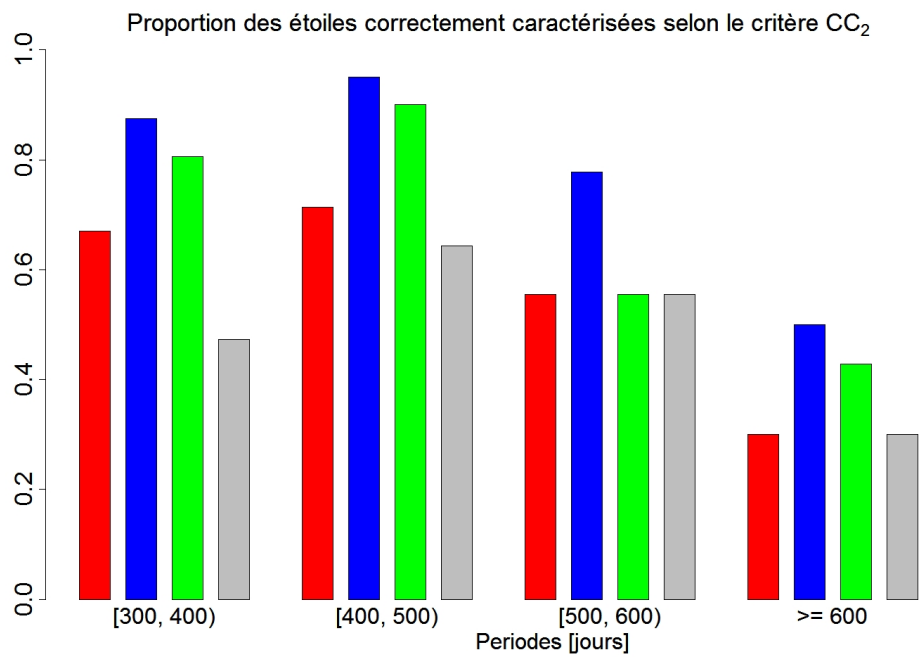
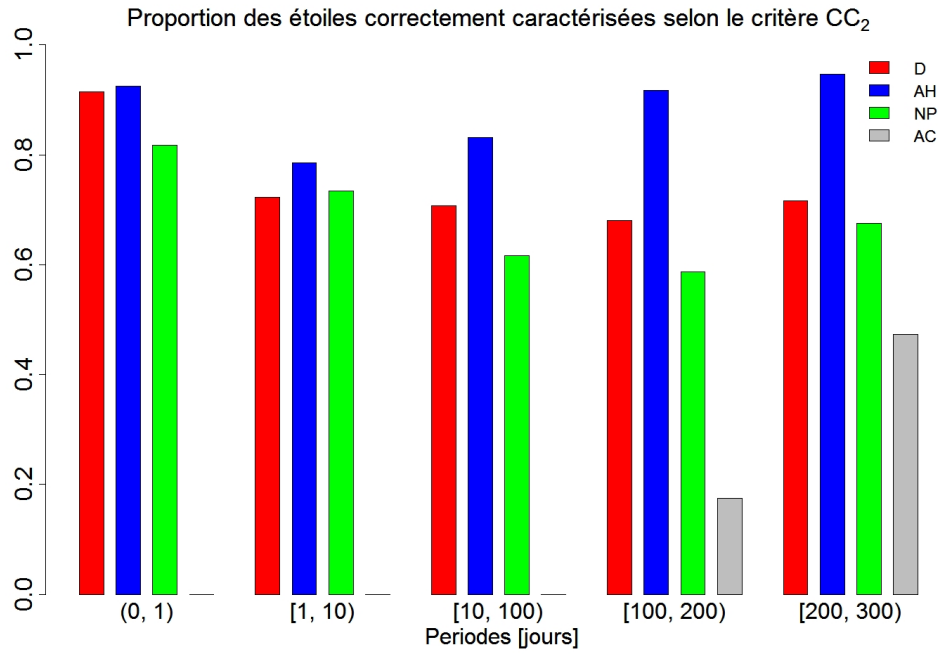


FIG. 3.3 – Graphiques représentant la proportion de caractérisations corrects des 2683 objets triés par période, selon le critère CC_2 .

783 objets ayant leur période comprise entre 0 et 1 jour.

		AH		Marges D			AH		Marges NP
		C	NC				C	NC	
D	C	694	22	716	NP	C	608	32	640
	NC	30	37			116	27	143	
Marges AH		724	59	783	Marges AH		724	59	783

1238 objets ayant leur période comprises entre 1 et 10 jours

		AH		Marges D			AH		Marges NP
		C	NC				C	NC	
D	C	844	50	894	NP	C	789	120	909
	NC	129	215			184	145	329	
Marges AH		973	265	1238	Marges AH		973	265	1238

297 objets ayant leur période comprises entre 10 et 100 jours

		AH		Marges D			AH		Marges NP
		C	NC				C	NC	
D	C	199	11	210	NP	C	167	16	183
	NC	48	39			80	34	114	
Marges AH		247	50	297	Marges AH		247	50	297

97 objets ayant leur période comprises entre 100 et 200 jours

		AH		Marges D			AH		Marges NP
		C	NC				C	NC	
D	C	60	6	66	NP	C	53	2	55
	NC	28	3			35	7	42	
Marges AH		88	9	97	Marges AH		88	9	97

95 objets ayant leur période comprises entre 200 et 300 jours

		AH		Marges D			AH		Marges NP
		C	NC				C	NC	
D	C	67	1	68	NP	C	61	4	65
	NC	22	5			28	2	30	
Marges AH		89	6	95	Marges AH		89	6	95

TAB. 3.4 – Le symbole C signifie correctement caractérisé et NC signifie faussement caractérisé. La ligne C représente les objets correctement caractérisés par la méthode D, pour le tableau de gauche et NP, pour le tableau de droite. La ligne NC représente les objets faussement caractérisés par la méthode D, pour le tableau de gauche et NP, pour le tableau de droite. Le nombre total d'objets C ou NC par D (ou NP) se lit dans la dernière colonne marges D (ou marges NP). La colonne C représente les objets correctement caractérisés par la méthode AH et la colonne NC représente les objets faussement caractérisés par la méthode AH, cela pour les deux tableaux. Le nombre total d'objets C ou NC par la méthode AH se lit dans la dernière ligne marges AH. La répartition du nombre d'objets C ou NC selon les méthodes donne un aperçu du nombre d'objets simultanément correctement caractérisés ou faussement caractérisés par les méthodes D, AH et NP.

112 objets ayant leur période comprises entre 300 et 400 jours

		AH		Marges D			AH		Marges NP
		C	NC				C	NC	
D	C	73	2	75	NP	C	82	8	90
	NC	25	12	37		NC	16	6	22
Marges AH		98	14	112	Marges AH		98	14	112

42 Objets ayant leur période comprises entre 400 et 500 jours

		AH		Marges D			AH		Marges NP
		C	NC				C	NC	
D	C	30	0	30	NP	C	36	0	36
	NC	8	4	12		NC	2	4	6
Marges AH		38	4	42	Marges AH		38	4	42

9 objets ayant leur période comprises entre 500 et 600 jours

		AH		Marges D			AH		Marges NP
		C	NC				C	NC	
D	C	4	1	5	NP	C	3	2	5
	NC	3	1	4		NC	4	0	4
Marges AH		7	2	9	Marges AH		7	2	9

10 objets ayant leur période comprises entre 600 jours et plus

		AH		Marges D			AH		Marges NP
		C	NC				C	NC	
D	C	3	0	3	NP	C	2	3	5
	NC	2	5	7		NC	3	2	5
Marges AH		5	5	10	Marges AH		5	5	10

TAB. 3.5 – Le symbole C signifie correctement caractérisé et NC signifie faussement caractérisé. La ligne C représente les objets correctement caractérisés par la méthode D, pour le tableau de gauche et NP, pour le tableau de droite. La ligne NC représente les objets faussement caractérisés par la méthode D, pour le tableau de gauche et NP, pour le tableau de droite. Le nombre total d'objets C ou NC par D (ou NP) se lit dans la dernière colonne marges D (ou marges NP). La colonne C représente les objets correctement caractérisés par la méthode AH et la colonne NC représente les objets faussement caractérisés par la méthode AH, cela pour les deux tableaux. Le nombre total d'objets C ou NC par la méthode AH se lit dans la dernière ligne marges AH. La répartition du nombre d'objets C ou NC selon les méthodes donne un aperçu du nombre d'objets simultanément correctement caractérisés ou faussement caractérisés par les méthodes D, AH et NP.

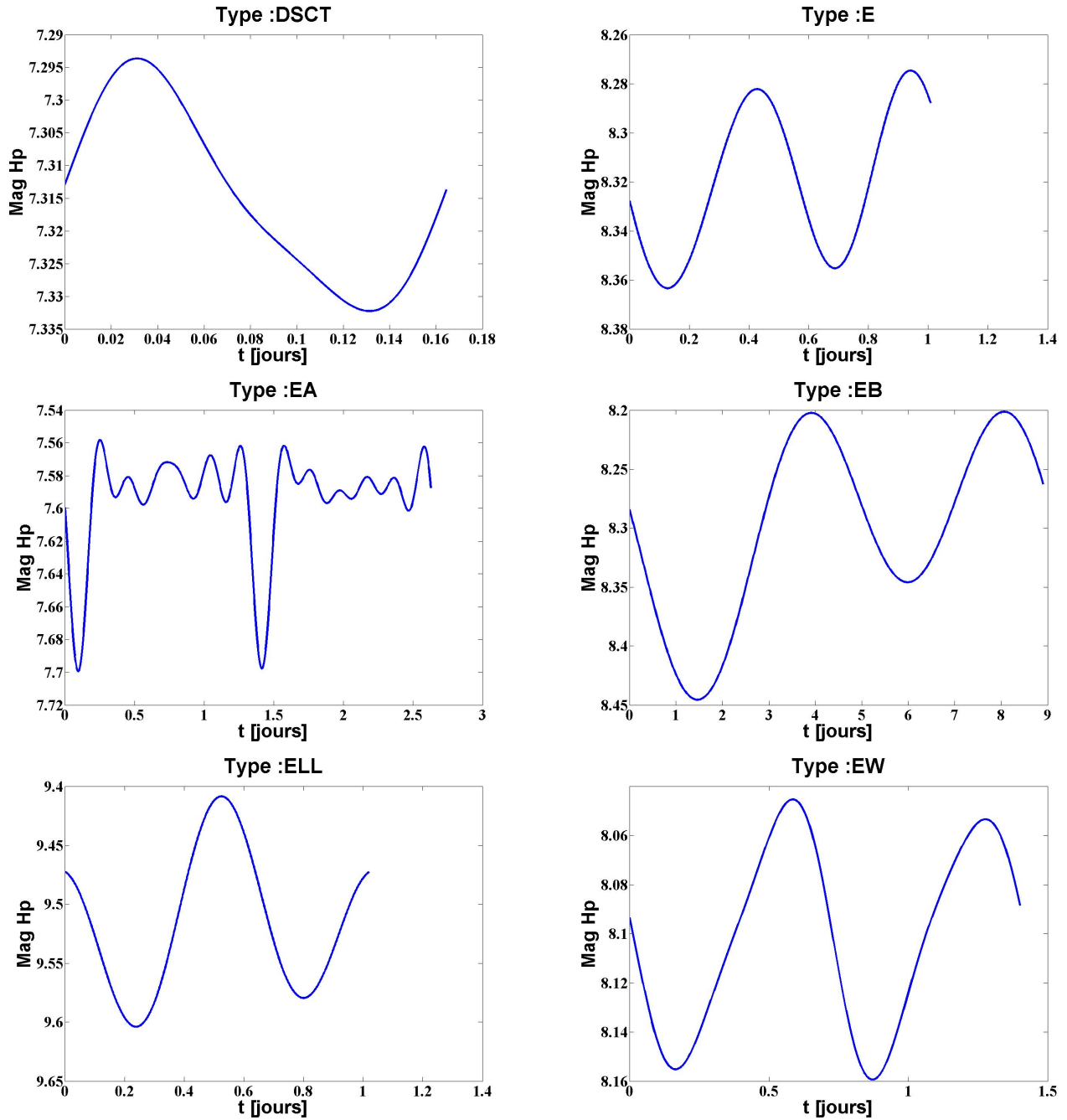


FIG. 3.4 – Séries de Fourier représentant les courbes lumières des objets de types DSCT, E, EA, EB, ELL et EW respectivement. La magnitude Mag est mesurée dans la bande photométrique Hp.

Type	Nombre d'objets de ce type
DSCT	85
E	20
EA	184
EB	240
ELL	31
EW	49
M	222
RRAB	145
RRC	34
SR	132

TAB. 3.6 – Objets triés par classes.

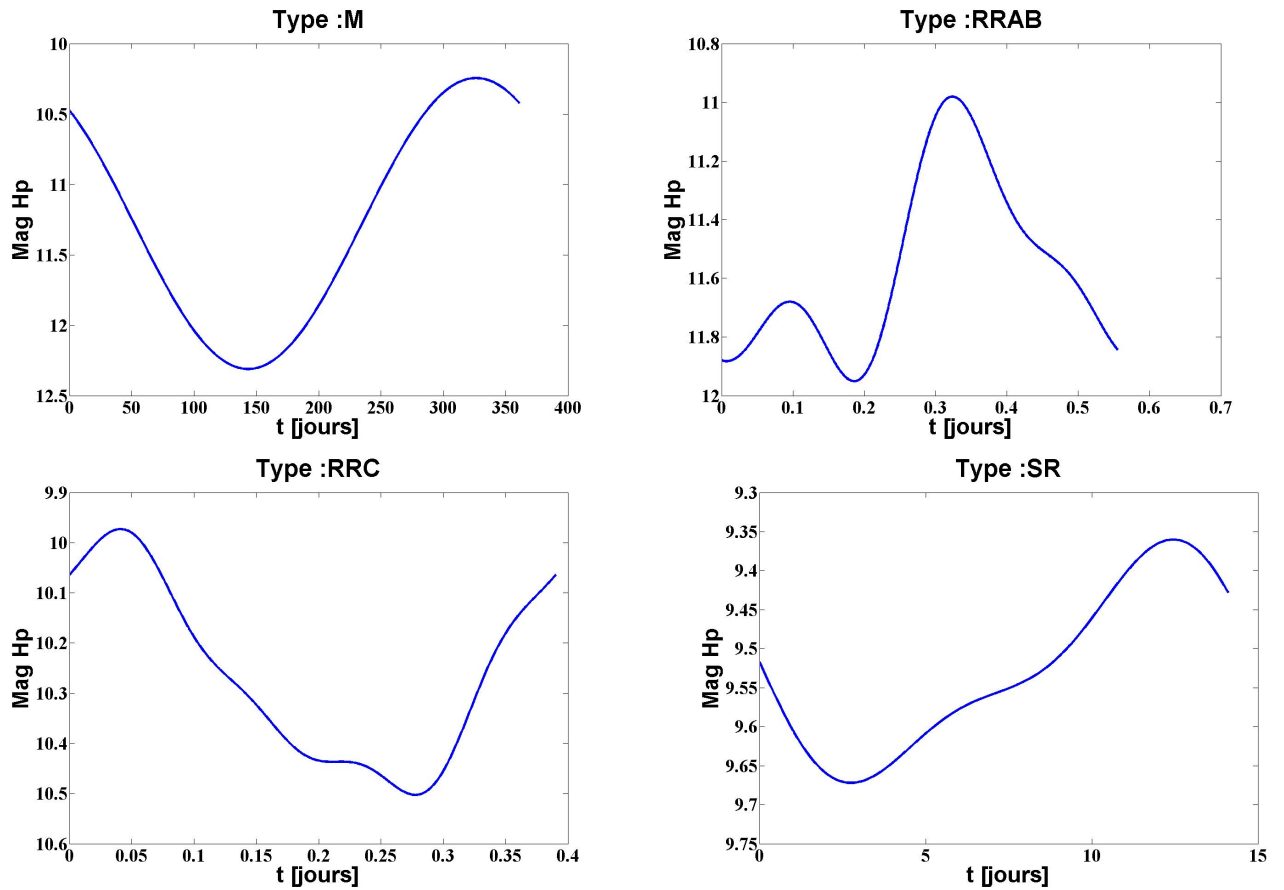


FIG. 3.5 – Séries de Fourier représentant les courbes lumières des objets de types M, RRAB, RRC et SR respectivement. La magnitude Mag est mesurée dans la bande photométrique Hp.

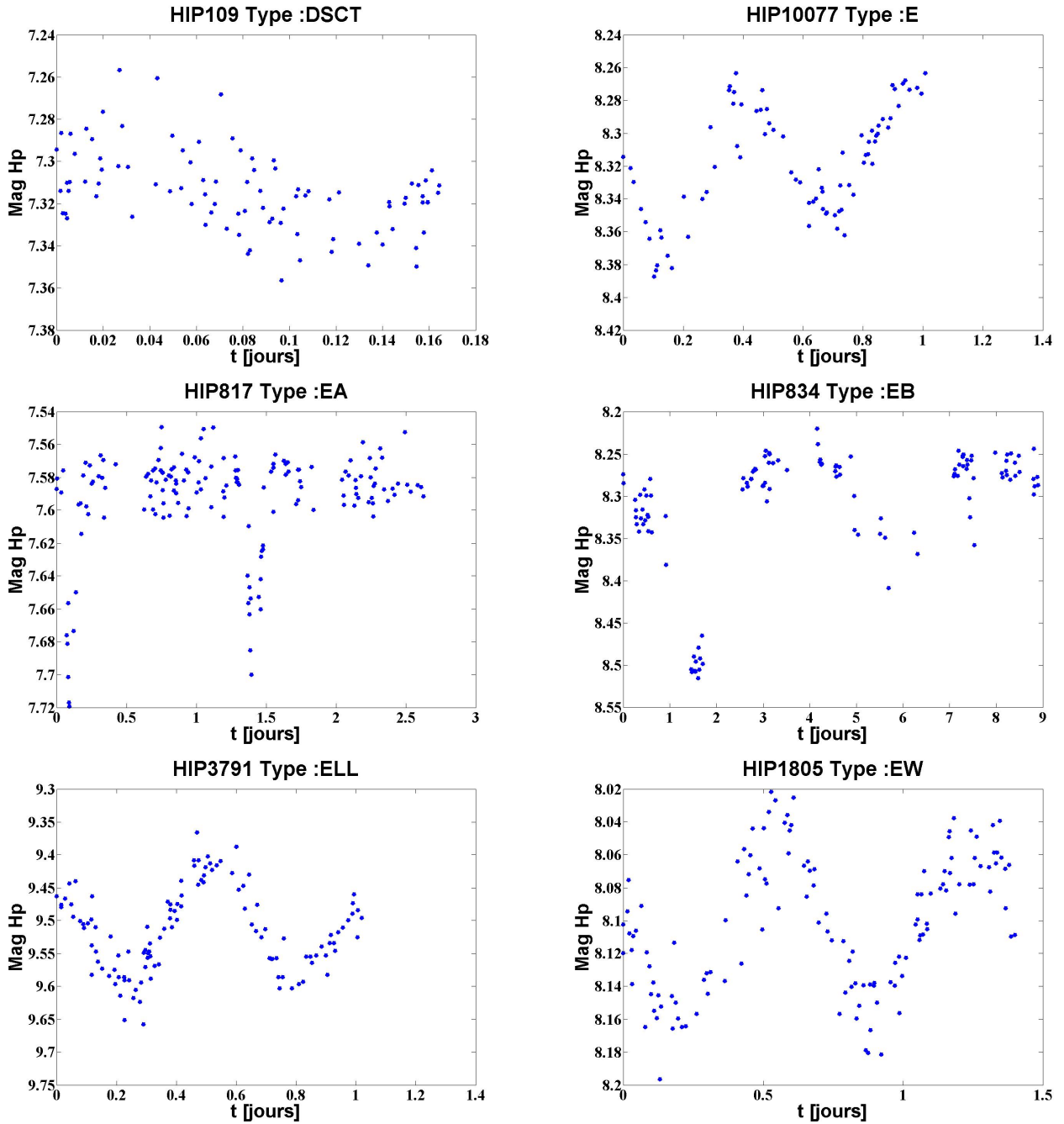


FIG. 3.6 – Observations de la courbe lumière repliées en phase des objets de types DSCT, E, EA, EB, ELL et EW respectivement. La magnitude Mag est mesurée dans la bande photométrique Hp.

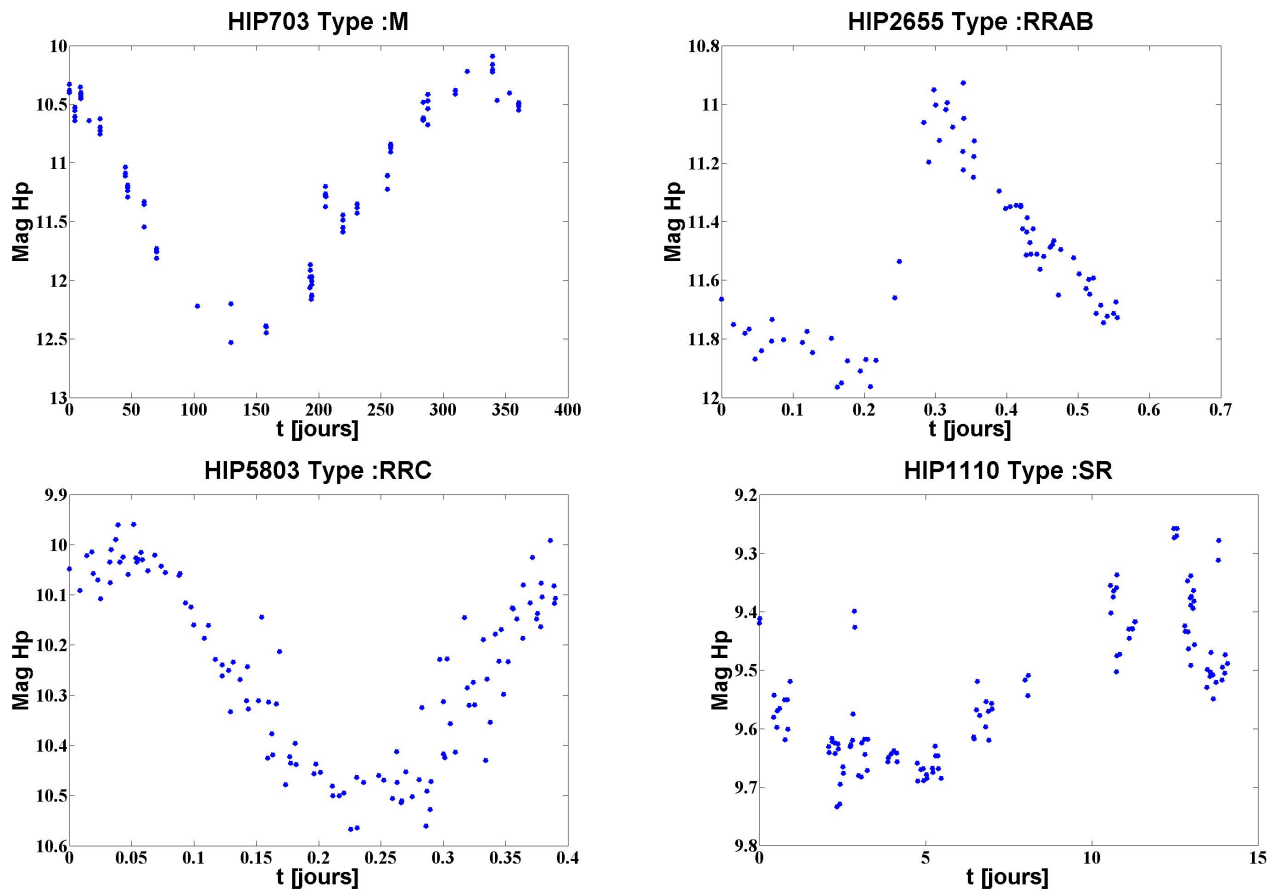


FIG. 3.7 – Observations de la courbe lumière repliées en phase des objets de types M, RRAB, RRC et SR respectivement. La magnitude Mag est mesurée dans la bande photométrique Hp.

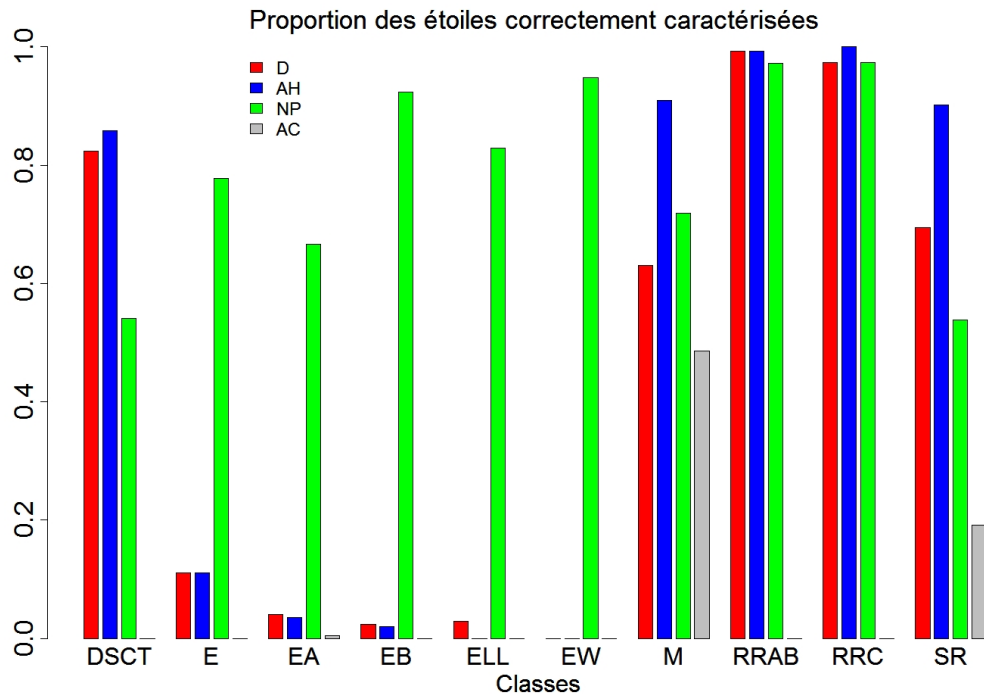


FIG. 3.8 – Graphiques représentant la proportion de caractérisations correctes de classes d’objets énumérés dans le tableau 3.6, selon le critère CC_1 . Les labels de l’axe des abscisses sont les classes des objets.

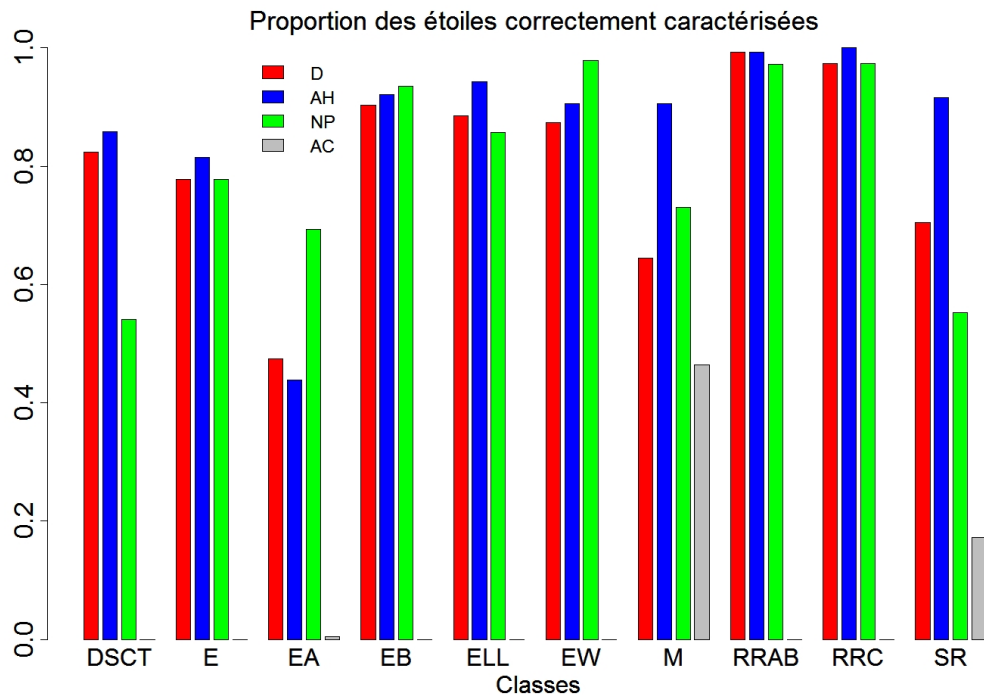


FIG. 3.9 – Graphiques représentant la proportion de caractérisations correctes de classes d’objets énumérés dans le tableau 3.6, selon le critère CC_2 . Les labels de l’axe des abscisses sont les classes des objets.

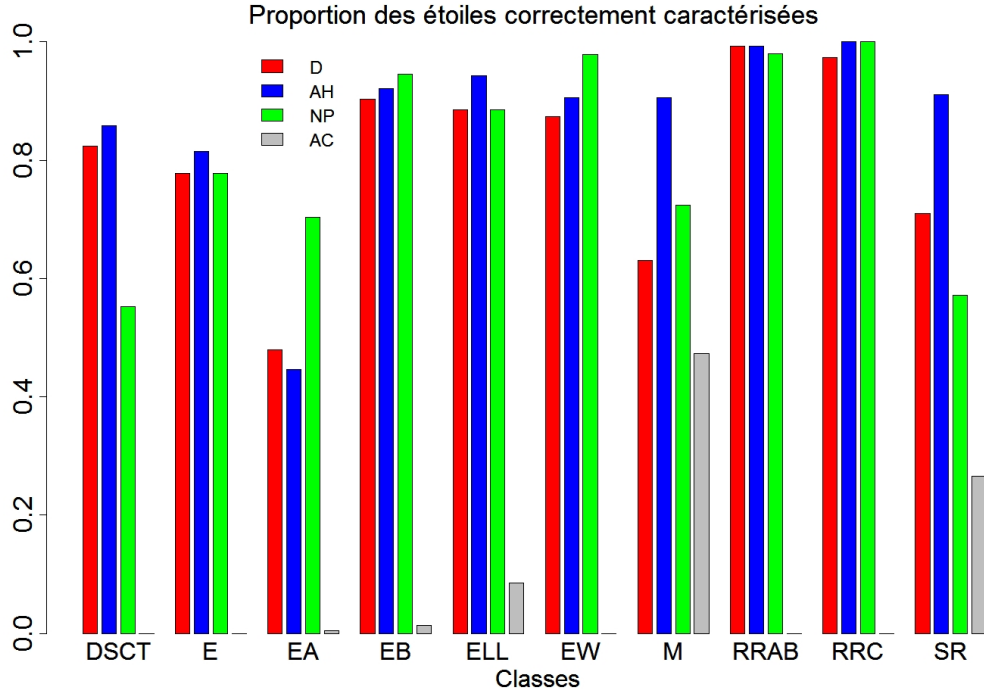


FIG. 3.10 – Graphiques représentant la proportion de caractérisation corrects de classes d’objets énumérés dans le tableau 3.6, selon le critère CC_3 . Les labels de l’axe des abscisse sont les classes des objets.

Dans les tableaux 3.7 et 3.8, on représente le nombre d’objets caractérisés (C) ou non caractérisés (NC) par les méthodes de Deeming (D) et de l’analyse harmonique (AH) (tableau de gauche), de la méthode de l’analyse harmonique et de la méthode non paramétrique (NP) (tableau de droite). A nouveau, la méthode AC n’a pas été considérée, au vu de sa trop mauvaise performance. On peut remarquer de très bonnes performances des méthodes D et AH sur les étoiles de type RRAB et RRC. Ces bons résultats montrent que, pour ces types d’objets, une régression sinusoidale avec la fréquence fondamentale fournit une bonne estimation de la fréquence.

3.2.4 Conclusion

L’application de la méthode de Deeming (D), de l’analyse harmonique (AH), de la régression non paramétrique (NP) et de la méthode basée sur l’autocorrélation (AC) aux données du catalogue A de la mission Hipparcos a permis de sélectionner la méthode AH comme étant la meilleure des quatre méthodes (voir la figure 3.2). A la même figure 3.2, on constate la très faible efficacité de la méthode AC. On observe aussi le problème de la multiplicité de la période pour la méthode NP, suivant le critère de comparaison considéré entre CC_1 et CC_3 . Le critère de comparaison CC_2 permet de détecter que les méthodes D et AH caractérisent des objets par leur demi-période au lieu de leur période. Contrairement à la méthode NP. Les figures 3.8 à 3.10 montrent que ce sont les objets de type binaires à éclipses (types E, EA, EB, ELL, EW) qui sont concernés. Ces types d’objets ont une courbe lumière de période P_0 , mais ressemblent fortement à un signal périodique de période $\frac{P_0}{2}$. On remarque aussi que la méthode NP ne souffre pas de ce problème. Nous avons aussi comparé les méthodes D et AH, ainsi que AH et NP, en considérant une fois les objets triés par périodes et une fois les objets triés par types d’objets. Nous avons analysé le nombre d’objets caractérisés en commun, ou seulement par l’une ou l’autre des méthodes. Les résultats montrent que les méthodes D et AH ont un grand nombre d’objets caractérisés en commun, mais la méthode AH a une performance supérieure. La comparaison entre les méthodes AH et NP montrent que la méthode AH est en général aussi supérieure à NP, sauf

85 étoiles de type DSCT

		AH		Marges D			AH		Marges NP
		C	NC				C	NC	
D	C	65	5	70	NP	C	43	4	47
	NC	8	7			30	8	38	
Marges AH		73	12	85	Marges AH		73	12	85

20 objets de type E

		AH		Marges D			AH		Marges NP
		C	NC				C	NC	
D	C	14	1	15	NP	C	13	1	14
	NC	1	4			5	2	4	
Marges AH		15	6	20	Marges AH		15	5	20

184 objets de type EA

		AH		Marges D			AH		Marges NP
		C	NC				C	NC	
D	C	87	14	101	NP	C	94	55	149
	NC	13	70			83	6	29	
Marges AH		100	84	184	Marges AH		100	84	184

240 objets de type EB

		AH		Marges D			AH		Marges NP
		C	NC				C	NC	
D	C	209	7	216	NP	C	213	16	229
	NC	11	13			24	7	4	
Marges AH		220	20	240	Marges AH		220	20	240

31 objets de type ELL

		AH		Marges D			AH		Marges NP
		C	NC				C	NC	
D	C	26	1	27	NP	C	26	1	27
	NC	3	1			4	3	1	
Marges AH		29	2	31	Marges AH		29	2	31

TAB. 3.7 – Le symbole C signifie correctement caractérisé et NC signifie faussement caractérisé. La ligne C représente les objets correctement caractérisés par la méthode D, pour le tableau de gauche et NP, pour le tableau de droite. La ligne NC représente les objets faussement caractérisés par la méthode D, pour le tableau de gauche et NP, pour le tableau de droite. Le nombre total d'objets C ou NC par D (ou NP) se lit dans la dernière colonne marges D (ou marges NP). La colonne C représente les objets correctement caractérisés par la méthode AH et la colonne NC représente les objets faussement caractérisés par la méthode AH, cela pour les deux tableaux. Le nombre total d'objets C ou NC par la méthode AH se lit dans la dernière ligne marges AH. La répartition du nombre d'objets C ou NC selon les méthodes donne un aperçu du nombre d'objets simultanément correctement caractérisés ou faussement caractérisés par les méthodes D, AH et NP.

49 objets de type EW

		AH		Marges D			AH		Marges NP
		C	NC				C	NC	
D	C	44	2	46	NP	C	44	4	48
	NC	1	2	3		NC	1	0	47
Marges AH		45	4	49	Marges AH		45	4	49

222 étoiles de type M

		AH		Marges D			AH		Marges NP
		C	NC				C	NC	
D	C	133	10	143	NP	C	189	18	207
	NC	68	11	79		NC	12	3	15
Marges AH		201	21	222	Marges AH		201	21	222

145 étoiles de type RRAB

		AH		Marges D			AH		Marges NP
		C	NC				C	NC	
D	C	144	0	144	NP	C	142	0	142
	NC	0	1	1		NC	2	1	3
Marges AH		144	1	145	Marges AH		144	1	145

34 étoiles de type RRC

		AH		Marges D			AH		Marges NP
		C	NC				C	NC	
D	C	33	0	33	NP	C	34	0	34
	NC	1	0	1		NC	0	0	0
Marges AH		34	0	34	Marges AH		34	0	34

132 étoiles de type SR

		AH		Marges D			AH		Marges NP
		C	NC				C	NC	
D	C	92	4	96	NP	C	80	1	81
	NC	30	6	36		NC	42	9	51
Marges AH		122	10	132	Marges AH		122	10	132

TAB. 3.8 – Le symbole C signifie correctement caractérisé et NC signifie faussement caractérisé. La ligne C représente les objets correctement caractérisés par la méthode D, pour le tableau de gauche et NP, pour le tableau de droite. La ligne NC représente les objets faussement caractérisés par la méthode D, pour le tableau de gauche et NP, pour le tableau de droite. Le nombre total d'objets C ou NC par D (ou NP) se lit dans la dernière colonne marges D (ou marges NP). La colonne C représente les objets correctement caractérisés par la méthode AH et la colonne NC représente les objets faussement caractérisés par la méthode AH, cela pour les deux tableaux. Le nombre total d'objets C ou NC par la méthode AH se lit dans la dernière ligne marges AH. La répartition du nombre d'objets C ou NC selon les méthodes donne un aperçu du nombre d'objets simultanément correctement caractérisés ou faussement caractérisés par les méthodes D, AH et NP.

dans le cas où l'on considère le cas particulier des binaires à éclipses. Nous avons aussi remarqué que la forme de la courbe lumière peut entraver la performance de la méthode AH. Dans les figures 3.8 à 3.10, on voit clairement une faible performance des méthodes D et AH sur les objets de type EA. La figure 3.6, affichant le repliement en phase d'un objet de type EA, montre que la forme de la fonction est très éloignée de la forme sinusoidale. De plus, la figure 3.4 montre une série de Fourier finie pour ce type d'objet. Un nombre élevé d'harmoniques est requis pour obtenir cette figure. La composante correspondant à la fréquence fondamentale ω_0 est faible, ce qui rend plus difficile la détection de cette fréquence. Cela montre la difficulté à caractériser la période d'un signal comportant une faible composante sinusoidale de fréquence ω_0 par les méthodes D et AH.

Chapitre 4

Distributions asymptotiques

Dans ce chapitre, nous allons en premier lieu parcourir l'histoire de la théorie asymptotique de l'estimateur de la fréquence. Par la suite, nous allons déterminer la distribution asymptotique dans le cas d'un échantillonnage irrégulier, vérifiant la condition de couverture de la période 2, d'un signal périodique. Nous donnerons aussi la variance asymptotique de l'estimateur de la fréquence pour des modèles ne vérifiant pas cette condition.

4.1 Historique de la théorie asymptotique

Le premier article donnant un résultat asymptotique de la distribution de l'estimateur de la fréquence est celui de Walker [32], en 1971. Il considère un modèle "signal + bruit", avec signal sinusoidal échantillonné de façon régulière. Le bruit est un processus Gaussien indépendant. Puis, Walker [33], en 1973, donne la distribution asymptotique de la fréquence pour le cas d'un signal sinusoidal échantillonné de manière régulière avec des erreurs corrélées. En 1981, Kovacs [15] propose une approximation de la variance asymptotique pour l'estimateur de la fréquence pour des observations régulières, qui est la même que celle trouvée par Walker [32]. En 1991, Schwarzenberg-Czerny [28] donne une variance asymptotique dans le cas d'observations irrégulières pour un cas particulier d'erreurs corrélées. En 2000, Hall et al. [13] proposent, ce que nous avons appelé la méthode non paramétrique pour une estimation simultanée de la forme du signal et de la période dans le cas d'observations irrégulières avec des erreurs indépendantes. Ils fournissent aussi la variance asymptotique de la période. Soit

$$Y_i = g_i + \varepsilon_i, i = 1, \dots, n, \quad (4.1)$$

où $g_i = g(t_i)$ est le signal continu $g(t)$, $t \in \mathbb{R}$ observé au temps t_i , ainsi que $\varepsilon_i = \varepsilon(t_i)$ est un processus Gaussien stationnaire et ergodique, observé au temps t_i , avec fonction d'autocovariance $\gamma(t)$. Ci-dessus, nous donnons les modèles pour la fonction $g(t)$ pour la méthode AH et NP, ainsi que les deux modèles pour le processus Gaussien stationnaires et ergodiques $\varepsilon(t)$. La méthode AH se base sur la décomposition $g(t) = s_g(t) + r_g(t)$ où $s_g(t) = A_g \cos(\omega_0 t) + B_g \sin(\omega_0 t) + \mu_g$ (voir section 2.1). Définissons la notation $g|_{[0, P_0)}$ comme étant la restriction de la fonction g à l'intervalle $[0, P_0)$. Nous supposons que la fonction g vérifie les conditions (4.2), (4.3) et (4.4) :

- g est une fonction continue, non constante, et périodique de période $P_0 > 0$; (4.2)
- $g|_{[0, P_0)}, g'|_{[0, P_0)}, g''|_{[0, P_0)} \in L^2([0, P_0))$, g', g'' sont la première et la deuxième dérivée de g ; (4.3)
- $A_g^2 + B_g^2 > 0$. (4.4)

De plus, nous supposons que la fonction d'autocovariance $\gamma(t)$ (1.4), du processus Gaussien $\varepsilon(t)$, vérifie la condition

$$\gamma \in L^2(\mathbb{R}). \quad (4.5)$$

Nous utiliserons les fonctions de Landau $O(\cdot)$, $o(\cdot)$, et leurs analogues stochastiques $O_p(\cdot)$ et $o_p(\cdot)$, $n \in \mathbb{N}$.

Définition 3 Soit une application $f : \mathbb{N} \rightarrow \mathbb{R}$.

- On appelle $f = O(n)$ s'il existe une constante $C \in \mathbb{R}$, et $n_0 \in \mathbb{N}$, tel que pour tout $n \geq n_0$, $|f(n)| < Cn$.
- De même, on appelle $f = o(n)$, si $\frac{f(n)}{n} \rightarrow 0$, lorsque $n \rightarrow \infty$.
- Soit $(Y_i)_{i=1}^n$, un processus stochastique. On appelle $Y_n = o_p(1)$ si pour tout $\varepsilon > 0$ et pour tout $\eta > 0$, il existe $n(\varepsilon, \eta) \in \mathbb{N}$ tel que si $n \geq n(\varepsilon, \eta)$, alors $\mathbb{P}[|Y_n| < \varepsilon] \geq 1 - \eta$. On appelle que $Y_n = o_p(n)$ si $\frac{Y_n}{n} = o_p(1)$, ou, de manière équivalente, $Y_n = n o_p(1)$.
- Soit $(Y_i)_{i=1}^n$, un processus stochastique. On appelle $Y_n = O_p(1)$ si pour tout $\eta > 0$, il existe une constante $K(\eta)$ et $n(\eta) \in \mathbb{N}$ tel que si $n \geq n(\eta)$, alors $\mathbb{P}[|Y_n| \leq K(\eta)] \geq 1 - \eta$. On appelle $Y_n = O_p(n)$ si $\frac{Y_n}{n} = O_p(1)$, ou, de manière équivalente, $Y_n = n O_p(1)$.

4.1.1 Signal sinusoidal, échantillonnage régulier et bruit Gaussien indépendant

Walker, [32], en 1971, a effectué la première étude de la distribution asymptotique de la fréquence pour n observations régulières d'un signal sinusoidal $s(t) = A \cos(\omega_0 t) + B \sin(\omega_0 t) + \mu$, de fréquence ω_0 et d'amplitude $\sqrt{A^2 + B^2} > 0$. Il considère le modèle suivant

$$Y_i = s(i) + \varepsilon_i, \quad (4.6)$$

où $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ iid. Il s'agit donc d'estimer la fonction de régression $s(t)$, c'est-à-dire de calculer les estimateurs $\hat{\omega}_n^{\text{AH}}$, \hat{A}_n , \hat{B}_n , et $\hat{\mu}_n$, définis aux points (2.16), (2.13), (2.14) et (2.15). On peut éliminer μ en considérant les données centrées $y_i^* = y_i - \bar{y}$. La fonction de régression devient donc $s(t) = A \cos(\omega t) + B \sin(\omega t)$. La somme des carrés des résidus du modèle réduit devient

$$\text{SC}(\omega, A, B) = \sum_{i=1}^n (y_i^* - s(i))^2. \quad (4.7)$$

Cet échantillonnage a une fréquence de Nyquist égale à π . Soit $\omega_k \in (0, \pi)$, une fréquence de Fourier. Alors les estimateurs des paramètres A et B valent

$$\hat{A}_n = \hat{A}_n(\omega_k) = \frac{\sum_{j=1}^n y_j^* \cos(\omega_k j)}{\sum_{j=1}^n \cos(\omega_k j)^2} = 2 \frac{\sum_{j=1}^n y_j^* \cos(\omega_k j)}{n}, \quad (4.8)$$

$$\hat{B}_n = \hat{B}_n(\omega_k) = \frac{\sum_{j=1}^n y_j^* \sin(\omega_k j)}{\sum_{j=1}^n \sin(\omega_k j)^2} = 2 \frac{\sum_{j=1}^n y_j^* \sin(\omega_k j)}{n}. \quad (4.9)$$

Pour une fréquence $\omega \in (0, \pi)$ quelconque, on trouve

$$\begin{aligned} \sum_{j=1}^n \cos(\omega j)^2 &= \frac{n}{2} + o_p(1), \\ \sum_{j=1}^n \sin(\omega j)^2 &= \frac{n}{2} + o_p(1), \\ \sum_{j=1}^n \sin(\omega j) \cos(\omega i) &= o_p(n). \end{aligned}$$

Si on considère les fréquences de Fourier $\omega = \omega_k = \frac{2\pi k}{n}$, alors

$$\begin{aligned}\sum_{j=1}^n \cos(\omega_k j)^2 &= \frac{n}{2}, \\ \sum_{j=1}^n \sin(\omega_k j)^2 &= \frac{n}{2}, \\ \sum_{j=1}^n \sin(\omega_k j) \cos(\omega_k j) &= 0.\end{aligned}$$

En considérant les équations (4.8) et (4.9), il en dérive les théorèmes suivants.

Théorème 4 (Walker, 1971) *Soit le modèle (4.6). Soit $\hat{\omega}_n^{AH}$ l'estimateur de la fréquence défini à l'équation (2.16). Alors, lorsque $n \rightarrow \infty$,*

$$\hat{\omega}_n^{AH} - \omega_0 = o_p(n^{-1}). \quad (4.10)$$

C'est-à-dire $n(\hat{\omega}_n^{AH} - \omega_0)$ converge en probabilité vers zéro, lorsque $n \rightarrow \infty$.

Ce théorème est donné sans preuve, car il est un cas particulier du théorème 8 qui sera démontré dans le cadre d'un signal g périodique de période P_0 , vérifiant les conditions (4.2) et (4.3), et avec des erreurs Gaussiennes corrélées.

Théorème 5 (Walker, 1971) *Soit le modèle (4.6). Soit $\hat{\omega}_n^{AH}$ l'estimateur de la fréquence défini à l'équation (2.16). Soit \hat{A}_n , \hat{B}_n et $\hat{\omega}_n^{AH}$ définis aux équations (4.8), (4.9) et (2.16) respectivement. Alors, lorsque $n \rightarrow \infty$, le vecteur $\left[n^{\frac{3}{2}}(\hat{\omega}_n^{AH} - \omega_0), n^{\frac{1}{2}}(\hat{A}_n - A), n^{\frac{1}{2}}(\hat{B}_n - B) \right]^T$ converge en distribution vers une loi Normale multivariée $\mathcal{N}(0, \sigma^2 W_{asy}^{-1})$, avec matrice de covariances $\sigma^2 W_{asy}^{-1}$ où W_{asy}^{-1} est l'inverse de*

$$W_{asy} = \begin{bmatrix} \frac{1}{6}(A^2 + B^2) & \frac{1}{4}B & -\frac{1}{4}A \\ \frac{1}{4}B & \frac{1}{2} & 0 \\ -\frac{1}{4}A & 0 & \frac{1}{2} \end{bmatrix}. \quad (4.11)$$

Ainsi la variance asymptotique de $\hat{\omega}_n^{AH}$ vaut $\sigma^2 (W_{asy}^{-1})_{11} = \frac{24\sigma^2}{(A^2+B^2)}$. Ce théorème est donné sans preuve, car il est un cas particulier du théorème 9 qui sera démontré dans le cadre d'un signal g périodique de période P_0 , vérifiant les conditions (4.2), (4.3) et (4.4), et d'erreurs Gaussiennes corrélées.

4.1.2 Signal sinusoidal, échantillonnage régulier et bruit Gaussien corrélé

En 1973, Walker [33] donne la distribution asymptotique de la fréquence pour le cas d'un signal sinusoidal observé de manière régulière, mais cette fois avec des erreurs corrélées. Il considère le modèle

$$Y_i = s(i) + \varepsilon_i, \quad i = 1, \dots, n \quad (4.12)$$

Le processus $(\varepsilon_i)_{i=1}^n$ est un processus AR(∞) de variance σ^2 vérifiant la relation $\varepsilon_i = \sum_{k=0}^{\infty} \psi_k \varepsilon_{i-k}$, $i \in \mathbb{N}$. Si les coefficients ψ_k sont tels que $\sum_{k=1}^{\infty} |\psi_k| < \infty$. Alors ε_i est un processus stochastique Gaussien, avec densité spectrale $f_{sp}(\omega) = \sigma^2 \left| \sum_{k=0}^{\infty} \psi_k \exp(i\omega k) \right|^2$. Walker énonce le théorème

Théorème 6 (Walker, 1973) *Soit le modèle (4.12). Le processus stochastique Gaussien a une densité spectrale $f_{sp}(\omega) = \sigma^2 \left| \sum_{k=0}^{\infty} \psi_k \exp(i\omega k) \right|^2$. Soit \hat{A}_n , \hat{B}_n et $\hat{\omega}_n^{AH}$, les estimateurs de paramètres A , B et ω_0 , définis aux équations (4.8), (4.9) et (2.16) respectivement. Alors, lorsque $n \rightarrow \infty$, le vecteur $\left[n^{\frac{3}{2}}(\hat{\omega}_n^{AH} - \omega_0), n^{\frac{1}{2}}(\hat{A}_n - A), n^{\frac{1}{2}}(\hat{B}_n - B) \right]^T$ converge en distribution vers une loi Normale multivariée $\mathcal{N}(0, f_{sp}(\omega_0) W_{asy}^{-1})$, avec matrice de covariances $f_{sp}(\omega_0) W_{asy}^{-1}$ où W_{asy}^{-1} est l'inverse de la matrice (4.11).*

Comme pour le théorème 5, ce théorème est donné sans preuve, car il est un cas particulier du théorème 9 qui sera démontré dans le cadre d'un signal g périodique de période P_0 , vérifiant les conditions (4.2), (4.3) et (4.4) et d'erreurs Gaussiennes corrélées.

4.2 Echantillonnage irrégulier vérifiant la condition de couverture de la période

4.2.1 Signal périodique quelconque, échantillonnage irrégulier et bruit Gaussien indépendant

En 2000, Hall et al. [13] proposent une méthode non paramétrique pour l'estimation simultanée de la forme de la fonction $g(t)$ et de la période P_0 . Ils démontrent le résultat suivant :

Théorème 7 (Hall et al., 2000) *Soit le modèle (4.1), avec erreur $\varepsilon(t)$, un processus stochastique Gaussien stationnaire et ergodique iid. La fonction $g(t)$ est périodique de période P_0 vérifiant les conditions (4.2) et (4.3). Soit $\hat{g}(t)$ l'estimateur de la fonction g , et $\hat{\omega}_n^{NP}$, l'estimateur de la fréquence ω_0 , définis à la section 2.4. L'échantillonnage est irrégulier, mais la condition de couverture de la période, définition 2, est vérifiée. De plus, $\eta = \lim_{i \rightarrow \infty} \mathbb{E}[T_i - T_{i-1}]$ existe. Alors, lorsque $n \rightarrow \infty$, $n^{\frac{3}{2}}(\hat{\omega}_n^{NP} - \omega_0)$ converge en distribution vers une loi Normale $\mathcal{N}(0, \tau^2)$, où*

$$\tau^2 = 24\pi\sigma^2\omega_0\eta^{-2} \left\{ \int_0^{P_0} g'(u)^2 du \right\}^{-1}. \quad (4.13)$$

Ce théorème est aussi donné sans démonstration, car il est un cas particulier du théorème 11, qui sera démontré dans le cadre d'erreurs corrélées.

4.2.2 Propriétés asymptotiques de l'estimateur $\hat{\omega}_n^{\text{AH}}$

Dans cette section, l'échantillonnage est irrégulier, avec la propriété de la couverture de l'intervalle de la période (voir chapitre 1.10 sur l'échantillonnage). Nous considérons une analyse harmonique (ou régression sinusoidale) sur des observations d'un signal périodique g quelconque de période P_0 , vérifiant les conditions (4.2), (4.3) et (4.4). Soit les paramètres $\omega_0, A_g, B_g, \mu_g$. Considérons le modèle (2.5), défini au chapitre 2 pour la méthode AH.

$$Y_i = g(t_i) + \varepsilon(t_i) = s_g(t_i) + r_g(t_i) + \varepsilon(t_i), \quad (4.14)$$

où $\varepsilon(t)$ est un processus stochastique en temps continu. L'estimateur minimise la somme des carrés (4.15).

$$\text{SC}(\omega, A, B, \mu) = \sum_{i=1}^n (Y_i - s(t_i))^2, \quad (4.15)$$

avec $s(t) = A \cos(\omega t) + B \sin(\omega t) + \mu$. On considère les observations centrées $y_i^* = y_i - \bar{y}$. La somme des carrés des résidus du modèle réduit devient

$$\text{SC}(\omega, A, B) = \sum_{i=1}^n (y_i^* - A \cos(\omega t_i) - B \sin(\omega t_i))^2. \quad (4.16)$$

La consistance de l'estimateur $\hat{\omega}_n^{\text{AH}}, \hat{A}_n, \hat{B}_n$, est démontré dans le premier paragraphe 4.2.2. Dans le second paragraphe 4.2.2, nous dérivons leurs distributions asymptotiques.

Consistence

Dans ce paragraphe, nous démontrons en particulier que $n(\widehat{\omega}_n^{AH} - \omega_0)$ converge en probabilité vers zéro, lorsque $n \rightarrow \infty$.

Théorème 8 (Ordre de convergence) *Soit le modèle (4.14), avec erreur $\varepsilon(t)$, un processus stochastique Gaussien stationnaire et ergodique corrélé, avec fonction d'autocovariance $\gamma(s)$ vérifiant (4.5). Supposons que la densité spectrale $f_{sp}(\omega)$ de l'erreur existe. La fonction $g(t)$ est périodique de période P_0 vérifiant les conditions (4.2), (4.3) et (4.4). Soit $\widehat{\omega}_n^{AH}$, l'estimateur de la fréquence ω_0 , \widehat{A}_n , l'estimateur de A_g , et \widehat{B}_n , l'estimateur de B_g , définis à la section 2.1. L'échantillonnage est irrégulier, mais vérifiant la définition 2 de la couverture de la période. De plus, $\eta = \lim_{i \rightarrow \infty} \mathbb{E}[T_i - T_{i-1}]$ existe. Alors, lorsque $n \rightarrow \infty$,*

$$\widehat{\omega}_n^{AH} - \omega_0 = o_p(n^{-1}), \quad (4.17)$$

$$\widehat{A}_n - A_g = o_p(1), \quad (4.18)$$

$$\widehat{B}_n - B_g = o_p(1). \quad (4.19)$$

La convergence est du même ordre que dans le cas d'erreurs indépendantes.

Preuve. Cette preuve suit la preuve donnée dans [32] pour le cas du signal sinusoidal avec observations régulières et erreurs indépendantes. On considère que les données sont centrées $y_k^* = y_k - \bar{y}$. Pour ne pas alourdir les calculs, on considère que la fonction g satisfait $\int_0^{P_0} g(t)dt = 0$. Considérons le périodogramme

$$I_n(\omega) = \frac{n}{2\pi} |J_n(\omega)|^2, \quad (4.20)$$

où

$$J_n(\omega) = \frac{1}{n} \sum_{k=1}^n y_k^* (\exp(i\omega t_k)) \quad (4.21)$$

est la transformée de Fourier discrète. On définit

$$\widehat{\omega}_n = \operatorname{argmax}_{\omega} I_n(\omega), \quad (4.22)$$

Premièrement nous allons démontrer que $(\widehat{\omega}_n - \omega_0) = o_p(n^{-1})$. Ensuite nous démontrerons que $(\widehat{\omega}_n^{AH} - \widehat{\omega}_n) = o_p(n^{-1})$. Développons la transformée discrète de Fourier,

$$\begin{aligned} J_n(\omega) &= \frac{1}{n} \sum_{k=1}^n (g(\tilde{t}_k) + \varepsilon_k) \exp(-i\omega t_k) \\ &= \frac{1}{n} \sum_{k=1}^n g(\tilde{t}_k) \exp(-i\omega t_k) + \frac{1}{n} \sum_{k=1}^n \varepsilon_k \exp(-i\omega t_k). \end{aligned} \quad (4.23)$$

Posons

$$J_{n,1}(\omega) = \frac{1}{n} \sum_{k=1}^n g(\tilde{t}_k) \exp(-i\omega t_k), \quad (4.24)$$

$$J_{n,2}(\omega) = \frac{1}{n} \sum_{k=1}^n \varepsilon_k \exp(-i\omega t_k). \quad (4.25)$$

Par la périodicité de la fonction $g(t)$ et $\exp(-i\omega t)$, lorsque $n \rightarrow \infty$, (4.24) s'écrit

$$\frac{1}{P_0} \int_0^{P_0} g(u) \exp(-i\omega u) du + O(n^{-1}) = \frac{1}{P_0} \mathcal{TF}\{g\}(\omega) + O(n^{-1}), \quad (4.26)$$

où $\mathcal{TF}\{g\}(\omega)$ est la transformée de Fourier de g . Ainsi le périodogramme $I_n(\omega)$ s'écrit

$$\begin{aligned} I_n(\omega) &= \frac{1}{2n\pi} \left[\left| \frac{1}{P_0} \mathcal{TF}\{g\}(\omega) \right|^2 + \right. \\ &\quad \left. 2\operatorname{Re} \left\{ \sum_{k=1}^n g(\tilde{t}_k) \exp(-i\omega \tilde{t}_k) \sum_{u=1}^n \varepsilon_u \exp(i\omega t_u) \right\} + \right. \\ &\quad \left. + \left| \sum_{k=1}^n \varepsilon_k \exp(i\omega t_k) \right|^2 \right], \end{aligned} \quad (4.27)$$

où on utilise le fait que $z + \bar{z} = 2\operatorname{Re}\{z\}$, $z \in \mathbb{C}$. Posons

$$I_{n,1}(\omega) = \left| \frac{1}{P_0} \mathcal{TF}\{g\}(\omega) \right|^2, \quad (4.28)$$

$$I_{n,2}(\omega) = \frac{1}{2n\pi} \left| \sum_{k=1}^n \varepsilon_k \exp(-i\omega \tilde{t}_k) \right|^2, \quad (4.29)$$

$$I_{n,3}(\omega) = \frac{1}{2n\pi} 2\operatorname{Re} \left\{ \sum_{k=1}^n g(\tilde{t}_k) \exp(-i\omega \tilde{t}_k) \sum_{u=1}^n \varepsilon_u \exp(i\omega t_u) \right\}. \quad (4.30)$$

Une des hypothèses du théorème sur la fonction g est $g|_{[0,P_0]} \in L^2([0,P_0])$, ce qui implique que $|\mathcal{TF}\{g\}(\omega)|^2 < M_g < \infty$, pour tout $\omega \in \mathbb{R}$. Lorsque $\omega = \omega_0$,

$$I_{n,1}(\omega_0) = \frac{n}{2\pi} \frac{1}{P_0^2} |\mathcal{TF}\{g\}(\omega_0)|^2 < \frac{n}{2\pi} \frac{1}{P_0^2} M_g = O(n). \quad (4.31)$$

Pour le calcul de $I_{n,2}(\omega)$, on considère son espérance $\mathbb{E}[I_{n,2}(\omega)] = \operatorname{Var}[J_{n,2}]$, car $\mathbb{E}[J_{n,2}(\omega)] = 0$. En effet,

$$\mathbb{E} \left[\frac{1}{n} \sum_{k=1}^n \varepsilon_k \exp(-i\omega t_k) \right] = \frac{1}{n} \sum_{k=1}^n \exp(-i\omega t_k) \mathbb{E}[\varepsilon_k] = 0,$$

car $\mathbb{E}[\varepsilon_k] = 0$, pour $k = 1, \dots, n$. Donc

$$\begin{aligned} \operatorname{Var} \left(\sum_{k=1}^n \varepsilon_k \exp(-i\omega t_k) \right) &= \sum_{k=1}^n \sum_{l=1}^n \exp(-i\omega t_k) \exp(i\omega t_l) \mathbb{E}[\varepsilon_k \varepsilon_l] \\ &= n\sigma^2 + 2 \sum_{k=1}^n \sum_{l=k+1}^n \exp(-i\omega(t_k - t_l)) \mathbb{E}[\varepsilon_k \varepsilon_l]. \end{aligned}$$

Considérons le module complexe de cette dernière expression, ainsi

$$\begin{aligned}
\left| n\sigma^2 + \sum_{k=1}^n \sum_{l=1}^n \exp(-i\omega(t_k - t_l)) \mathbb{E}[|t_k - t_l|] \right| &\leq n\sigma^2 + 2 \sum_{k=1}^n \sum_{l=k+1}^n \mathbb{E}[|\varepsilon_k \varepsilon_l|] \\
&\leq n\sigma^2 + 2 \sum_{k=1}^n \mathbb{E} \left[\sqrt{\left(\sum_{l=k+1}^n \varepsilon_k \varepsilon_l \right)^2} \right] \\
&\leq n\sigma^2 + 2 \sum_{k=1}^n \sqrt{\mathbb{E} \left[\left(\sum_{l=k+1}^n \varepsilon_k \varepsilon_l \right)^2 \right]} \\
&\leq n\sigma^2 + 2 \sum_{k=1}^n \sqrt{\mathbb{E} \left[\sum_{l=k+1}^n \sum_{m=k+1}^n \varepsilon_k \varepsilon_l \varepsilon_k \varepsilon_m \right]} \\
&\leq n\sigma^2 + 2 \sum_{k=1}^n \sqrt{\mathbb{E} \left[\varepsilon_k^2 \sum_{l=k+1}^n \sum_{m=k+1}^n \varepsilon_l \varepsilon_m \right]}. \quad (4.32)
\end{aligned}$$

Calculons l'espérance de l'équation (4.32). En premier lieu remarquons que

$$\begin{aligned}
\left(\sum_{l=k+1}^n \varepsilon_l \right)^2 &= \sum_{l=k+1}^n \sum_{m=k+1}^n \varepsilon_l \varepsilon_m \\
&= \sum_{l=k+1}^n \varepsilon_l^2 + 2 \sum_{l=k+1}^n \sum_{m>l}^n \varepsilon_l \varepsilon_m \\
&= \sum_{l=k+1}^n \varepsilon_l^2 + 2 \sum_{l=k+1}^n \sum_{m=l+1}^n \varepsilon_l \varepsilon_m.
\end{aligned}$$

Ainsi,

$$\mathbb{E} \left[\varepsilon_k^2 \sum_{l=k+1}^n \sum_{m=k+1}^n \varepsilon_l \varepsilon_m \right] = \sum_{l=k+1}^n \mathbb{E} [\varepsilon_k^2 \varepsilon_l^2] + 2 \sum_{l=k+1}^n \sum_{m=l+1}^n \mathbb{E} [\varepsilon_k^2 \varepsilon_l \varepsilon_m]. \quad (4.33)$$

Nous nous occupons en premier lieu de la première somme de (4.33). Remarquons que

$$\sum_{l=k+1}^n \mathbb{E} [\varepsilon_k^2 \varepsilon_l^2] = \sum_{l=k+1}^n \text{Var} [\varepsilon_k \varepsilon_l] + \mathbb{E} [\varepsilon_k \varepsilon_l]^2. \quad (4.34)$$

En conditionnant $\varepsilon_k \varepsilon_l$ avec ε_k , et considérant la formule de la variance $\mathbb{E} [\text{Var} [\varepsilon_k \varepsilon_l | \varepsilon_k]] + \text{Var} [\mathbb{E} [\varepsilon_k \varepsilon_l | \varepsilon_k]]$, (4.34) devient

$$\sum_{l=k+1}^n \mathbb{E} [\text{Var} [\varepsilon_k \varepsilon_l | \varepsilon_k]] + \text{Var} [\mathbb{E} [\varepsilon_k \varepsilon_l | \varepsilon_k]] + \gamma(|t_k - t_l|)^2. \quad (4.35)$$

Avec le conditionnement sur ε_k , on peut écrire (4.35) comme

$$\sum_{l=k+1}^n \mathbb{E} [\varepsilon_k \text{Var} [\varepsilon_l | \varepsilon_k]] + \text{Var} [\varepsilon_k \mathbb{E} [\varepsilon_l | \varepsilon_k]] + \gamma(|t_k - t_l|)^2. \quad (4.36)$$

En considérant que $\varepsilon_k \sim \mathcal{N}(0, \sigma^2)$, $\varepsilon_l | \varepsilon_k = e_k \sim \mathcal{N}\left(\frac{\gamma(|t_k - t_l|)}{\sigma^2} e_k, \sigma^2 - \frac{\gamma(|t_k - t_l|)^2}{\sigma^2}\right)$ et que $\mathbb{E}(\varepsilon_k^4) = 3\sigma^4$, (4.36) devient

$$\begin{aligned} \sum_{l=k+1}^n \text{Var} \left[\varepsilon_k \frac{\gamma(|t_k - t_l|)}{\sigma^2} \varepsilon_k \right] + \gamma(|t_k - t_l|)^2 &= \sum_{l=k+1}^n \frac{\gamma(|t_k - t_l|)^2}{\sigma^4} \text{Var}[\varepsilon_k^2] + \gamma(|t_k - t_l|)^2 \\ &= \sum_{l=k+1}^n \frac{\gamma(|t_k - t_l|)^2}{\sigma^4} 2\sigma^4 + \gamma(|t_k - t_l|)^2 \\ &= 3 \sum_{l=k+1}^n \gamma(|t_k - t_l|)^2. \end{aligned} \quad (4.37)$$

Maintenant, considérons la relation (1.5) entre la densité spectrale $f_{sp}(\omega)$ du processus Gaussien et la fonction d'autocovariance $\gamma(t)$. En remplaçant $\gamma(|t_k - t_l|)$ par $\frac{1}{2\pi} \int_{-\infty}^{\infty} f_{sp}(\omega) \exp(-i\omega |t_l - t_k|) d\omega$, avec $t_l > t_k$, dans (4.37), on obtient

$$\begin{aligned} 3 \frac{1}{4\pi^2} \sum_{l=k+1}^n \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{sp}(\omega) f_{sp}(\xi) \exp(i\omega(t_l - t_k)) \exp(i\xi(t_l - t_k)) d\omega d\xi \\ = 3 \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{sp}(\omega) f_{sp}(\xi) \sum_{l=k+1}^n \exp(i(\omega + \xi)(t_l - t_k)) d\omega d\xi. \end{aligned} \quad (4.38)$$

Remarquons que, lorsque $n \rightarrow \infty$, $t_n \simeq n$, on peut écrire $\sum_{l=k+1}^n \exp(-i(\omega + \xi)(t_l - t_k)) \simeq D_n^+(\omega + \xi)$, où $D_n^+(\omega)$ est la version complexe (car la somme des temps ne se fait que de 0 à n , et non de $-n$ à n) du noyau de Dirichlet (1.29). Pour éviter cela, on peut approximer $\sum_{l=k+1}^n \gamma(|t_k - t_l|)$ par $\frac{1}{2} \sum_{l=-n+2k+1}^n \gamma(|t_k - t_l|)$, grâce à la symétrie de $\gamma(x)$ par rapport à zéro. Ainsi, en effectuant la même procédure, on obtient $\frac{1}{2} \sum_{l=-n+2k+1}^n \exp(-i(\omega + \xi)(t_l - t_k)) \simeq D_n(\omega + \xi)$. Le noyau de Dirichlet (1.29) tend vers la fonction de Dirac $\delta(\omega + \xi)$, lorsque $n \rightarrow \infty$. On peut donc approximer (4.38) par

$$\frac{3}{4\pi^2} \frac{1}{2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{sp}(\omega) f_{sp}(\xi) \delta(\omega + \xi) d\omega d\xi + O(n^{-1}),$$

qui est égal à

$$\frac{3}{8\pi^2} \int_{-\infty}^{\infty} f_{sp}(\omega) f_{sp}(-\omega) d\omega + O(n^{-1}) = \frac{3}{8\pi^2} \int_{-\infty}^{\infty} f_{sp}(\omega)^2 d\omega + O(n^{-1}) < \frac{3}{8\pi^2} M_{f_{sp}}. \quad (4.39)$$

A la dernière ligne (4.39), on a utilisé la symétrie de la densité spectrale $f_{sp}(\omega) = f_{sp}(-\omega)$ et que $f_{sp} \in L^2[\mathbb{R}]$, car $\gamma \in L^2[\mathbb{R}]$. Pour la deuxième somme de (4.33), nous utilisons la formule de la covariance partielle

$$\text{Cov}(X, Y | Z) = \text{Cov}(X, Y) - \frac{\text{Cov}(Z, X) \text{Cov}(X, Y)}{\text{Cov}(Z, Z)}, \quad (4.40)$$

avec $X = \varepsilon_l$, $Y = \varepsilon_m$, $Z = \varepsilon_k$. Nous obtenons

$$\begin{aligned}
2 \sum_{l=k+1}^n \sum_{m=l+1}^n \mathbb{E} [\varepsilon_k^2 \varepsilon_l \varepsilon_m] &= 2 \mathbb{E} \left[\sum_{l=k+1}^n \sum_{m=l+1}^n \mathbb{E} [\varepsilon_k^2 \varepsilon_l \varepsilon_m \mid \varepsilon_k] \right] \\
&= 2 \mathbb{E} \left[\sum_{l=k+1}^n \sum_{m=l+1}^n \varepsilon_k^2 \mathbb{E} [\varepsilon_l \varepsilon_m \mid \varepsilon_k] \right] \\
&= 2 \mathbb{E} \left[\sum_{l=k+1}^n \sum_{m=l+1}^n \varepsilon_k^2 \left(\gamma(|t_l - t_m|) - \frac{\gamma(|t_l - t_k|)\gamma(|t_l - t_m|)}{\sigma^2} + \frac{\varepsilon_k^2 \gamma(|t_l - t_k|)\gamma(|t_m - t_k|)}{\sigma^4} \right) \right] \\
&= 2 \sum_{l=k+1}^n \sum_{m=l+1}^n (\sigma^2 \gamma(|t_l - t_m|) - \gamma(|t_l - t_k|)\gamma(|t_l - t_m|) \\
&\quad + 3\gamma(|t_l - t_k|)\gamma(|t_m - t_k|)). \tag{4.41}
\end{aligned}$$

Premièrement, considérons le premier terme de la somme (4.41). La somme $\sum_{l=k+1}^n \sum_{m=l+1}^n \gamma(|t_l - t_m|)$ contient $\frac{(n-k)(n-k-1)}{2}$ termes, et l'intervalle d'observation est de longueur $\sim n - k$, lorsque $n \rightarrow \infty$. On peut approximer $\sum_{l=k+1}^n \sum_{m=l+1}^n \gamma(|t_l - t_m|)$ par $\frac{(n-k)(n-k-1)}{n-k} \sum_{i=1}^n \sum_{j=1}^n \gamma(|t_i - t_j|) \frac{n}{(n-k)\frac{n}{2}} \simeq \frac{n-k}{2} \int_0^n \gamma(x) dx$. Ainsi

$$\begin{aligned}
\sum_{l=k+1}^n \sum_{m=l+1}^n \gamma(|t_l - t_m|) &= \frac{n-k}{2} \int_0^n \gamma(x) dx + O((n-k)^{-1}) \\
&= \frac{n-k}{4} f_{sp}(0) + O((n-k)^{-1}). \tag{4.42}
\end{aligned}$$

Pour le deuxième terme, considérons à nouveau la relation (1.4) entre la densité spectrale $f_{sp}(\omega)$ du processus Gaussien et la fonction d'autocovariance $\gamma(t)$. Alors $\gamma(|t_l - t_k|) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f_{sp}(\omega) \exp(-i\omega(t_l - t_k)) d\omega$, avec $t_l > t_k$, et $\gamma(|t_l - t_m|) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f_{sp}(\xi) \exp(-i\xi(t_m - t_l)) d\xi$, avec $t_m > t_l$. De plus, approximations $\sum_{l=k+1}^n \sum_{m=l+1}^n \gamma(|t_l - t_k|)\gamma(|t_l - t_m|)$ par $\frac{1}{4} \sum_{l=-n+2k+1}^n \sum_{m=-n+2l+1}^n \gamma(|t_l - t_k|)\gamma(|t_l - t_m|)$. Ainsi

$$\frac{1}{16\pi^2} \sum_{l=-n+2k+1}^n \sum_{m=-n+2l+1}^n \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{sp}(\omega) f_{sp}(\xi) \exp(-i\omega(t_l - t_k)) \exp(-i\xi(t_m - t_l)) d\omega d\xi,$$

où l'argument de l'intégrale double peut se réécrire comme

$$f_{sp}(\omega) f_{sp}(\xi) \exp(i(\omega t_k)) \exp(-i(\omega - \xi)t_l) \exp(-i\xi t_m).$$

Lorsque $n \rightarrow \infty$, $t_n \simeq n$. De plus, remplaçons $\sum_{l=-n+2k+1}^n \exp(-i(\omega - \xi)t_l)$ par $D_n((\omega - \xi)) + O(n^{-1})$ ainsi que $\sum_{m=-n+2l+1}^n \exp(-i\xi t_m)$ par $D_n(\xi) + O(n^{-1})$, où $D_n(\xi)$ est le noyau de Dirichlet (1.29). De plus, lorsque $n \rightarrow \infty$, $D_n(\xi) \rightarrow \delta(\xi)$ et $D_n(\omega - \xi) \rightarrow \delta(\omega - \xi)$, la fonction de Dirac. On peut donc écrire (4.43), lorsque $n \rightarrow \infty$, comme

$$(n-k) \frac{1}{4\pi^2} \int_{-\infty}^{\infty} f_{sp}(\omega)^2 \delta(\omega) d\omega + O(n^{-1}) = (n-k) \frac{1}{4\pi^2} f_{sp}(0)^2 + O(n^{-1}). \tag{4.43}$$

Pour le troisième terme, en considérant à nouveau la densité spectrale comme pour le deuxième terme, nous obtenons que $\sum_{l=k+1}^n \sum_{m=l+1}^n \gamma(|t_l - t_k|)\gamma(|t_m - t_k|)$ s'écrit

$$\frac{1}{16\pi^2} \sum_{l=-n+2k+1}^n \sum_{m=-n+2l+1}^n \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{sp}(\omega) f_{sp}(\xi) \exp(i\omega(t_l - t_k)) \exp(i\xi(t_m - t_k)) d\omega d\xi. \quad (4.44)$$

On peut écrire $D_n(\xi) = \sum_{m=-n+2k+1}^n \exp(-i\xi(t_m - t_k)) + O(n^{-1})$ et $D_n(\omega) = \sum_{l=-n+2k+1}^n \exp(-i\omega(t_l - t_k)) + O(n^{-1})$, où $D_n(\omega)$ est le noyau de Dirichlet (1.29). De plus, lorsque $n \rightarrow \infty$, $D_n(\omega) \rightarrow \delta(\omega)$, la fonction de Dirac. On peut écrire (4.44) comme

$$\frac{1}{16\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{sp}(\omega) f_{sp}(\xi) \delta(\omega) \delta(\xi) d\omega d\xi + O(n^{-1}) = \frac{1}{16\pi^2} f_{sp}(0)^2 + O(n^{-1}). \quad (4.45)$$

Nous reprenons donc le point (4.32) avec les résultats (4.42) à (4.45). De plus, nous définissons $h_{sp}(\omega)$ comme $f_{sp}(\omega) = \sigma^2 h_{sp}(\omega)$. C'est-à-dire que $h_{sp}(0) = 2 \int_0^{+\infty} \frac{\gamma(x)}{\sigma^2} dx = 2 \int_0^{+\infty} \rho(x) dx$, où $\rho(x)$ est l'autocorrélation du processus Gaussien. Nous avons donc

$$\begin{aligned} n\sigma^2 + 2 \sum_{k=1}^n \sqrt{\sum_{l=k+1}^n \sum_{m=k+1}^n \mathbb{E} [\varepsilon_k^2 \varepsilon_l \varepsilon_m]} \\ \leq n\sigma^2 + 2 \sum_{k=1}^n \sqrt{\left(\frac{(n-k)}{2} f_{sp}(0) \sigma^2 - (n-k) \frac{f_{sp}(0)^2}{16\pi^2} + \frac{3}{16\pi^2} f_{sp}(0)^2 \right)} \\ = n\sigma^2 + 2\sigma^2 \sum_{k=1}^n \sqrt{(n-k) \left(\frac{1}{2} h_{sp}(0) - \frac{1}{16\pi^2} h_{sp}(0)^2 \right) + \frac{3}{16\pi^2} h_{sp}(0)^2} \\ < n\sigma^2 + 2\sigma^2 \sqrt{h_{sp}(0)} \sum_{k=1}^n \sqrt{\frac{1}{2}(n-k) + \frac{3}{16\pi^2} h_{sp}(0)} \\ = n\sigma^2 + 2\sigma^2 \sqrt{h_{sp}(0)} \sum_{k=1}^n \sqrt{\frac{1}{2}(n-k) + O(n)} \\ = n\sigma^2 + \sigma^2 \sqrt{h_{sp}(0)} \frac{4}{3} \left(\frac{1}{2} n \right)^{\frac{3}{2}} + O(n) \\ = n\sigma^2 \left(1 + \sqrt{h_{sp}(0)} \frac{1}{6} \frac{n^{\frac{3}{2}}}{n} \right) + O(n) \\ = n\sigma^2 \left(1 + \sqrt{h_{sp}(0)} \frac{1}{6} \sqrt{n} \right) + O(n) \\ < C_0 n^{\frac{3}{2}} \sqrt{h_{sp}(0)} \sigma^2, \end{aligned} \quad (4.46)$$

où $C_0 > 0$ est une constante tel que $C_0 n^{\frac{3}{2}} \sqrt{h_{sp}(0)} > n \left(1 + \sqrt{h_{sp}(0)} \frac{1}{6} \sqrt{n} \right)$, pour $n > 1$. Ainsi

$$\left| \sum_{k=1}^n \varepsilon_k \exp(i\omega t_k) \right|^2 = O_p(n^{\frac{3}{2}}). \quad (4.47)$$

Et pour le terme mixte, on a

$$I_{n,3}(\omega) = O(n) O_p(n^{\frac{3}{4}}) = O_p(n^{\frac{7}{4}}). \quad (4.48)$$

Posons

$$K_n(\delta) = \max_{0 \leq \omega \leq \pi} |I_n(\omega)|.$$

Par (4.27), (4.31), (4.47) et (4.48), on a que

$$\max_{0 \leq \omega \leq \pi} \left| I_n(\omega) - \frac{1}{n} \left| \sum_{k=1}^n g(t_k) \exp(-i\omega t_k) \right|^2 \right| = O_p(n^{\frac{3}{4}}) + O_p(n^{\frac{1}{2}}). \quad (4.49)$$

Pour montrer que

$$\lim_{n \rightarrow \infty} \mathbb{P}[n|\omega - \omega_0| \geq \delta] = 0, \quad (4.50)$$

on peut montrer que

$$\lim_{n \rightarrow \infty} \mathbb{P}[K_n(\delta) \geq I_n(\omega_0)] = 0. \quad (4.51)$$

Nous avons vu, au point (4.31), que lorsque $n \rightarrow \infty$,

$$I_{n,1}(\omega_0) < \lim_{n \rightarrow \infty} \frac{n}{2\pi} \frac{1}{P_0^2} M_g. \quad (4.52)$$

Alors

$$\begin{aligned} K_n(\delta) &= \max_{|\omega - \omega_0| \leq \frac{\delta}{n}} |I_n(\omega)| \\ &\leq I_{n,1}(\omega_0) + O_p(n^{\frac{7}{4}}) \\ &< \frac{n}{2\pi} \frac{1}{P_0^2} M_g + O_p(n^{\frac{7}{4}}). \end{aligned} \quad (4.53)$$

$$(4.54)$$

Ainsi

$$\frac{1}{n} K_n(\delta) < \frac{1}{2\pi P_0^2} M_g = \lim_{n \rightarrow \infty} \frac{\mathbb{P}}{n} I_n(\omega_0). \quad (4.55)$$

Ce qui veut dire que

$$\lim_{n \rightarrow \infty} \mathbb{P}[K_n(\delta) < I_n(\omega_0)] = 1. \quad (4.56)$$

C'est-à-dire que si $|\omega - \omega_0| \geq \frac{\delta}{n}$, alors on a le résultat (4.56) et donc ceci implique les résultats (4.51) et (4.50). Maintenant nous démontrons que $(\hat{\omega}_n^{AH} - \hat{\omega}_n) = o_p(n^{-1})$. Pour cela, remarquons que $\hat{\omega}_n$, $\tilde{A}_n = 2 \frac{\sum_{k=1}^n y_k^{star} \cos(\omega t_k)}{n}$ et $\tilde{B}_n = 2 \frac{\sum_{i=1}^n y_i^* \sin(\omega t_i)}{n}$ minimisent la quantité

$$Q(\theta) = \sum_{k=1}^n (y_k^*)^2 - 2 \sum_{i=1}^n y_i (A \cos(\omega t_i) + B \sin(\omega t_i)) + \frac{1}{2} n (A^2 + B^2), \quad (4.57)$$

(on peut voir cela dans [32]). Montrons que $\tilde{A}_n - A_{g,1} = o_p(1)$ et $\tilde{B}_n - B_{g,1} = o_p(1)$.

$$\tilde{A}_n = \frac{n}{2} \sum_{k=1}^n g(t_k) \cos(\hat{\omega}_n t_k) + \frac{n}{2} \sum_{k=1}^n \varepsilon_k \cos(\hat{\omega}_n t_k), \quad (4.58)$$

$$\tilde{B}_n = \frac{n}{2} \sum_{k=1}^n g(t_k) \sin(\hat{\omega}_n t_k) + \frac{n}{2} \sum_{k=1}^n \varepsilon_k \sin(\hat{\omega}_n t_k). \quad (4.59)$$

$$(4.60)$$

Par le modèle (4.14), on a que

$$g(t) = s_g(t) + r_g(t). \quad (4.61)$$

Par le fait que $\hat{\omega}_n$ converge vers ω_0 lorsque $n \rightarrow \infty$, le lemme 2 de la démonstration du théorème 9 montre que

$$\begin{aligned} \sum_{k=1}^n r_g(t_k) \cos(\omega_0 t_k) &= o(1), \\ \sum_{k=1}^n r_g(t_k) \sin(\omega_0 t_k) &= o(1). \end{aligned}$$

Ainsi, lorsque $n \rightarrow \infty$,

$$\begin{aligned} \tilde{A}_n &= \frac{2}{n} \sum_{k=1}^n s_g(t_k) \cos(\hat{\omega}_n t_k) + \frac{2}{n} \sum_{k=1}^n \varepsilon_k \cos(\hat{\omega}_n t_k) = \frac{2}{n} A_g \sum_{k=1}^n \cos(\omega_0 t_k)^2 + \frac{2}{n} \sum_{k=1}^n \varepsilon_k \cos(\hat{\omega}_n t_k) + o(1), \\ \tilde{B}_n &= \frac{2}{n} \sum_{k=1}^n s_g(t_k) \sin(\hat{\omega}_n t_k) + \frac{2}{n} \sum_{k=1}^n \varepsilon_k \sin(\hat{\omega}_n t_k) = \frac{2}{n} B_g \sum_{k=1}^n \sin(\omega_0 t_k)^2 + \frac{2}{n} \sum_{k=1}^n \varepsilon_k \sin(\hat{\omega}_n t_k) + o(1). \end{aligned}$$

Remarquons que

$$\begin{aligned} \sum_{k=1}^n \cos(\omega_0 t_k)^2 &= \sum_{k=1}^n \left(\frac{1}{2} + \cos(2\omega_0 t_k) \right) = \frac{n}{2} + O(1) = \frac{n}{2} (1 + 2o(1)), \\ \sum_{k=1}^n \sin(\omega_0 t_k)^2 &= \sum_{k=1}^n \left(\frac{1}{2} - \cos(2\omega_0 t_k) \right) = \frac{n}{2} + O(1) = \frac{n}{2} (1 + 2o(1)). \end{aligned}$$

De plus, avec le point (4.47), $\sum_{k=1}^n \varepsilon_k \sin(\hat{\omega}_n t_k) = O_p(n^{\frac{3}{4}})$, nous obtenons

$$\tilde{A}_n = A_{g,1} \frac{2}{n} \left(\frac{n}{2} (1 + o(1)) \right) + \frac{2}{n} O_p(n^{\frac{3}{4}}) = A_g + o_p(1), \quad (4.62)$$

$$\tilde{B}_n = B_{g,1} \frac{2}{n} \left(\frac{n}{2} (1 + o(1)) \right) + \frac{2}{n} O_p(n^{\frac{3}{4}}) = B_g + o_p(1). \quad (4.63)$$

Ainsi, \tilde{A}_n et \tilde{B}_n convergent vers A_g et B_g en $o_p(1)$. L'estimateur $\hat{\theta}_n = [\hat{\omega}_n^{\text{AH}} \hat{A}_n \hat{B}_n]$ du paramètre $\theta_g = [\omega_0 \ A_g \ B_g]$ minimise la fonction des moindres carrés

$$\text{SC}(\theta) = \sum_{k=1}^n (y_k^* - s_g(t_k))^2. \quad (4.64)$$

où

$$\begin{aligned}\widehat{A}_n &= \frac{\sum_{k=1}^n y_k^* \cos(\widehat{\omega}_n^{\text{AH}} t_k)}{\sum_{k=1}^n \cos(\widehat{\omega}_n^{\text{AH}} t_k)^2}, \\ \widehat{B}_n &= \frac{\sum_{k=1}^n y_k^* \sin(\widehat{\omega}_n^{\text{AH}} t_k)}{\sum_{k=1}^n \sin(\widehat{\omega}_n^{\text{AH}} t_k)^2}.\end{aligned}$$

Nous avons vu plus haut que

$$\begin{aligned}\sum_{k=1}^n \cos(\widehat{\omega}_n^{\text{AH}} t_k)^2 &= \frac{n}{2}(1 + 2o(1)), \\ \sum_{k=1}^n \sin(\widehat{\omega}_n^{\text{AH}} t_k)^2 &= \frac{n}{2}(1 + 2o(1)).\end{aligned}$$

Ainsi, en utilisant le développement limité de degré 1 en 0 de $\frac{1}{1+x} = 1 + x + o(n^{-1})$ autour de $x = 0$, on peut écrire \widehat{A}_n et \widehat{B}_n comme

$$\begin{aligned}\widehat{A}_n &= \frac{\sum_{k=1}^n y_k^* \cos(\widehat{\omega}_n^{\text{AH}} t_k)}{\frac{n}{2}(1 + o(1))} = \frac{\sum_{k=1}^n y_k^* \cos(\widehat{\omega}_n^{\text{AH}} t_k)}{\frac{n}{2}}(1 + o(1)) = \widetilde{A}_n + o(1), \\ \widehat{B}_n &= \frac{\sum_{k=1}^n y_k^* \sin(\widehat{\omega}_n^{\text{AH}} t_k)}{\frac{n}{2}(1 + o(1))} = \frac{\sum_{k=1}^n y_k^* \sin(\widehat{\omega}_n^{\text{AH}} t_k)}{\frac{n}{2}}(1 + o(1)) = \widetilde{B}_n + o(1).\end{aligned}$$

Pour pouvoir déterminer l'ordre de convergence de $\widehat{\omega}_n^{\text{AH}} - \widetilde{\omega}_n$, on approxime (4.64) par un développement limité d'ordre 2 dans un voisinage du paramètre $\widetilde{\theta}_n$. Ceci donne

$$\text{SC}(\widehat{\theta}_n) = \text{SC}(\widetilde{\theta}_n) + \nabla \text{SC}(\widetilde{\theta}_n)(\widehat{\theta}_n - \widetilde{\theta}_n) + \frac{1}{2}(\widehat{\theta}_n - \widetilde{\theta}_n)^t H(\text{SC}(\widetilde{\theta}_n))(\widehat{\theta}_n - \widetilde{\theta}_n) + O_p(n^{-1}), \quad (4.65)$$

où $H(\text{SC}(\widetilde{\theta}_n))$ est la matrice Hessienne de $\text{SC}(\widetilde{\theta}_n)$. En prenant la première dérivée à gauche et à droite de l'égalité (4.65), nous obtenons

$$\nabla \text{SC}(\widehat{\theta}_n) = H(\text{SC}(\widetilde{\theta}_n))(\widehat{\theta}_n - \widetilde{\theta}_n) + o_p(n^{-1}).$$

Ceci revient à

$$(\widehat{\theta}_n - \widetilde{\theta}_n) = \left(-H(\text{SC}(\widetilde{\theta}_n))\right)^{-1} \left(\widetilde{\nabla} \text{SC}(\widetilde{\theta}_n) + o_p(n^{-1})\right). \quad (4.66)$$

Maintenant, nous pouvons calculer les ordres de la matrice $H(\text{SC}(\widetilde{\theta}_n))$. Par le lemme 4 (preuve du théorème 9), nous avons que

$$H(\text{SC}(\widetilde{\theta}_n)) = \begin{bmatrix} O(n^3) & O(n^2) & O(n^2) \\ O(n^2) & O(n) & O(1) \\ O(n^2) & O(1) & O(n) \end{bmatrix}. \quad (4.67)$$

La matrice inverse $\left(-H(\text{SC}(\widetilde{\theta}_n))\right)^{-1} = \frac{1}{\det(H(\text{SC}(\widetilde{\theta}_n)))} \text{Cof}\left(H(\text{SC}(\widetilde{\theta}_n))\right)^t$, où $\text{Cof}\left(H(\text{SC}(\widetilde{\theta}_n))\right)$ est matrice des cofacteurs de $H(\text{SC}(\widetilde{\theta}_n))$ et $\det\left(H(\text{SC}(\widetilde{\theta}_n))\right)$ est le déterminant de $H(\text{SC}(\widetilde{\theta}_n))$. Le déterminant est d'ordre

$$\det\left(H(\text{SC}(\widetilde{\theta}_n))\right) = O(n^3)O(n^2) = O(n^5),$$

et les composantes de la matrice des cofacteurs sont d'ordres

$$\text{Cof}\left(H(\text{SC}(\tilde{\theta}_n))\right) = \begin{bmatrix} O(n^2) & O(n^3) & O(n^3) \\ O(n^3) & O(n^4) & O(n^4) \\ O(n^3) & O(n^4) & O(n^4) \end{bmatrix}. \quad (4.68)$$

Ainsi, l'ordre des composantes de la matrice inverse $\left(-H(\text{SC}(\tilde{\theta}_n))\right)^{-1}$ sont

$$\left(-H(\text{SC}(\tilde{\theta}_n))\right)^{-1} = - \begin{bmatrix} O(n^{-3}) & O(n^{-2}) & O(n^{-2}) \\ O(n^{-2}) & O(n^{-1}) & O(n^{-1}) \\ O(n^{-2}) & O(n^{-1}) & O(n^{-1}) \end{bmatrix}. \quad (4.69)$$

Il nous manque encore les ordres de $\left(\nabla\text{SC}(\tilde{\theta}_n) + o_p(n^{-1})\right)$.

$$\nabla\text{SC}(\tilde{\theta}_n) = -2 \sum_{k=1}^n (y_k^* - s_{\tilde{\theta}_n}(t_k)) \begin{bmatrix} \tilde{B}_n \cos(\hat{\omega}_n t_k) t_k - \tilde{A}_n \sin(\hat{\omega}_n t_k) t_k \\ \cos(\hat{\omega}_n t_k) \\ \sin(\hat{\omega}_n t_k) \end{bmatrix}. \quad (4.70)$$

Par les résultats (4.62), (4.63) et par les équations (4.65) et (4.65), nous avons que \hat{A}_n converge vers \tilde{A}_n et que \tilde{A}_n converge vers A_0 . De mêmes pour \hat{B}_n et \tilde{A}_n . Ainsi, par le modèle (4.14), on a que

$$y_k^* - s_{\tilde{\theta}_n}(t_k) = r_{\theta_g}(t_k) + \varepsilon(t_k) + o(1), \quad (4.71)$$

car $\tilde{\theta}_n$ converge vers θ_g . Par le lemme 2 de la preuve du théorème 9, on a que $\sum_{i=1}^n r_{\tilde{\theta}_n}(t_k) \nabla s_{\tilde{\theta}_n}(t_k) = o(1)$, où

$$\nabla s_{\tilde{\theta}_n}(t_k) = \begin{bmatrix} \tilde{B}_n \cos(\hat{\omega}_n t_k) t_k - \tilde{A}_n \sin(\hat{\omega}_n t_k) t_k \\ \cos(\hat{\omega}_n t_k) \\ \sin(\hat{\omega}_n t_k) \end{bmatrix}.$$

Ainsi (4.70) peut s'écrire

$$-2 \sum_{k=1}^n \varepsilon(t_k) \nabla s_{\tilde{\theta}_n}(t_k) + o(1). \quad (4.72)$$

Calculons maintenant les ordres des composantes de (4.72). La première composante devient

$$\sum_{k=1}^n \varepsilon(t_k) \left(\tilde{B}_n \cos(\hat{\omega}_n t_k) t_k - \tilde{A}_n \sin(\hat{\omega}_n t_k) t_k \right) = \tilde{B}_n \frac{\partial}{\partial \omega} \sum_{k=1}^n \varepsilon_k \cos(t_k) + \tilde{A}_n \frac{\partial}{\partial \omega} \sum_{k=1}^n \varepsilon_k \sin(t_k). \quad (4.73)$$

Nous pouvons donc écrire $\frac{\partial}{\partial \omega} \sum_{k=1}^n \varepsilon_k \cos(\omega t_{(k)}) = o_p(n)$ et $\frac{\partial}{\partial \omega} \sum_{k=1}^n \varepsilon_k \sin(\omega t_{(k)}) = o_p(n)$. Ainsi la première composante de (4.72) est d'ordre $o_p(n)$. Pour la deuxième et la troisième composante de (4.72), on obtient un ordre de $o_p(n)$, car, comme pour la première composante, $\sum_{k=1}^n \varepsilon(t_k) \cos(\hat{\omega}_n t_k) = o_p(n)$ et $\sum_{k=1}^n \varepsilon(t_k) \sin(\hat{\omega}_n t_k) = o_p(n)$. Ainsi

$$\left(\nabla\text{SC}(\tilde{\theta}_n) + o_p(n^{-1})\right) = \begin{bmatrix} o_p(n) \\ o_p(n) \\ o_p(n) \end{bmatrix}. \quad (4.74)$$

Donc, par les résultats ci-dessus, nous obtenons

$$(\widehat{\theta}_n - \widetilde{\theta}_n) = \left(-H(\text{SC}(\widetilde{\theta}_n))\right)^{-1} \left(\vec{\nabla}\text{SC}(\widetilde{\theta}_n) + o_p(n^{-1})\right) = \begin{bmatrix} o_p(n^{-1}) \\ o_p(1) \\ o_p(1) \end{bmatrix}. \quad (4.75)$$

La première composante de (4.75) montre que $(\widehat{\omega}_n^{\text{AH}} - \widetilde{\omega}_n) = o_p(n^{-1})$. La deuxième et troisième composante montre que $\widehat{A}_n - \widetilde{A}_n = o_p(1)$ et $\widehat{B}_n - \widetilde{B}_n = o_p(1)$. ■

Distribution asymptotique

Nous démontrons que $\left[n^{\frac{3}{2}}(\widehat{\omega}_n^{\text{AH}} - \omega_0), n^{\frac{1}{2}}(\widehat{A}_n - A_g), n^{\frac{1}{2}}(\widehat{B}_n - B_g)\right]^T$ a une distribution asymptotique normale d'espérance nulle et de variance $f_{sp}(\omega_0)W_{asy}^{-1}$, où W_{asy}^{-1} est l'inverse de la matrice (4.105) du théorème 5, et $f_{sp}(\omega)$ est la densité spectrale du processus gaussien corrélé.

Théorème 9 (Distribution asymptotique des estimateurs $\widehat{\omega}_n^{\text{AH}}, \widehat{A}_n, \widehat{B}_n$) Soit le modèle (4.14), avec erreur $\varepsilon(t)$, un processus stochastique Gaussien stationnaire et ergodique corrélé, avec fonction d'autocovariance $\gamma(s)$ vérifiant (4.5). Supposons que la densité spectrale $f_{sp}(\omega)$ de l'erreur existe. La fonction $g(t)$ est périodique de période P_0 vérifiant les conditions (4.2), (4.3) et (4.4). Soit $\widehat{\omega}_n^{\text{AH}}$, l'estimateur de la fréquence ω_0 , défini à la section 2.1. L'échantillonnage est irrégulier, mais vérifiant la définition 2 de la couverture de la période. De plus, $\eta = \lim_{i \rightarrow \infty} \mathbb{E}[T_i - T_{i-1}]$ existe. Alors, lorsque $n \rightarrow \infty$, $\left[n^{\frac{3}{2}}(\widehat{\omega}_n^{\text{AH}} - \omega_0), n^{\frac{1}{2}}(\widehat{A}_n - A_g), n^{\frac{1}{2}}(\widehat{B}_n - B_g)\right]^T$ converge en distribution vers une loi Normale multivariée $\mathcal{N}(0, f_{sp}(\omega_0)W_{asy}^{-1})$, où la matrice de covariances W_{asy}^{-1} est l'inverse de la matrice (4.11). En particulier, la variance asymptotique de $n^{\frac{3}{2}}\widehat{\omega}_n^{\text{AH}}$ vaut

$$\text{Var}_{asy} \left[n^{\frac{3}{2}}(\widehat{\omega}_n^{\text{AH}}) \right] = \frac{24f_{sp}(\omega_0)}{A_g^2 + B_g^2}. \quad (4.76)$$

Preuve. Les estimateurs $\widehat{\omega}_n^{\text{AH}}, \widehat{A}_n$ et \widehat{B}_n minimisent la fonction des moindres carrés

$$\text{SC}(\omega, A, B) = \sum_{i=1}^n (Y_i - A \cos(\omega t_i) - B \sin(\omega t_i))^2. \quad (4.77)$$

Pour pouvoir déterminer le comportement des estimateurs dans un voisinage des vrais paramètres inconnus $\theta_g \in \mathbb{R}^4$, on approxime (4.77) par un développement limité d'ordre 2 dans un voisinage des vrais paramètres. Avec la notation $\theta = (\omega, A, B)^T \in \mathbb{R}^3$ et $\theta_g = (\omega_0, A_g, B_g)^T \in \mathbb{R}^3$ ceci donne

$$\text{SC}(\theta) = \text{SC}(\theta_g) + \nabla\text{SC}(\theta_g)(\theta - \theta_g) + \frac{1}{2}(\theta - \theta_g)^t H(\text{SC}(\theta_g))(\theta - \theta_g) + o_p(n^{-1}), \quad (4.78)$$

où $\nabla\text{SC}(\theta)$, évalué en θ_g , est le gradient de $\text{SC}(\theta)$ et $H(\text{SC}(\theta))$, évalué en θ_g , est la matrice Hessienne de $\text{SC}(\theta)$. En prenant la première dérivée à gauche et à droite de l'égalité (4.78), nous obtenons

$$\nabla\text{SC}(\theta_g) = H(\text{SC}(\theta_g))(\theta_g - \theta) + o_p(n^{-1}).$$

En considérant $\theta = \widehat{\theta}_n$, ceci revient à

$$(\widehat{\theta}_n - \theta_g) = (-H(\text{SC}(\theta_g)))^{-1} \left(\vec{\nabla}\text{SC}(\theta_g) + o_p(n^{-1})\right). \quad (4.79)$$

Pour nous permettre de calculer la loi limite, lorsque $n \rightarrow \infty$, de $(\hat{\theta}_n - \theta)$, réécrivons la ligne (4.79) comme

$$\sqrt{n}(\hat{\theta}_n - \theta_g) = \left(-\frac{1}{n} \text{H}(\text{SC}(\theta_g)) \right)^{-1} \left(\frac{1}{\sqrt{n}} \nabla \text{SC}(\theta_g) + o_p(n^{-\frac{1}{2}}) \right), \quad (4.80)$$

La loi de $\sqrt{n}(\hat{\theta}_n - \theta_g)$ tend vers une loi normale $\mathcal{N}(\mu_\theta, \Sigma_\theta)$. Posons $s_g(t) = A \cos(\omega t) + B \sin(\omega t)$. Ainsi,

$$\nabla \text{SC}(\theta_g) = \sum_{i=1}^n (y_i - s_g(t_i)) \nabla s_g(t_i).$$

Sous l'hypothèse de notre modèle, pour $n \rightarrow \infty$, $(y_i - s_g(t_i)) \sim \mathcal{N}(r_g(t_i), \sigma^2)$. Ainsi $\sum_{i=1}^n (y_i - s_g(t_i)) \nabla s_g(t_i)$ est une somme de vecteur de variables aléatoires normales dont la résultante est une variable aléatoire normale avec moyenne $\mu_{\theta_g, n} = \sum_{i=1}^n r_g(t_i) \nabla s_g(t_i)$ et avec une certaine variance Σ_{θ_g} .

$$\nabla \text{SC}(\theta_g) = - \sum_{i=1}^n (y_i - s_g(t_i)) \begin{bmatrix} B_g \cos(\omega_0 t_i) t_i - A_g \sin(\omega_0 t_i) t_i \\ \cos(\omega_0 t_i) \\ \sin(\omega_0 t_i) \end{bmatrix}, \quad (4.81)$$

où

$$\nabla s_g(t_i) = \begin{bmatrix} B_g \cos(\omega_0 t_i) t_i - A_g \sin(\omega_0 t_i) t_i \\ \cos(\omega_0 t_i) \\ \sin(\omega_0 t_i) \end{bmatrix}. \quad (4.82)$$

Lemme 2 Soit $g(t) \in L^2(\mathbb{R})$, une fonction périodique de période P_0 , vérifiant les conditions (4.2), (4.3) et (4.4). L'échantillonnage est irrégulier, mais vérifiant la définition 2 de la couverture de la période. De plus, $\eta = \lim_{i \rightarrow \infty} \mathbb{E}[T_i - T_{i-1}]$ existe. Posons $\mu_{\theta_g, n} = \sum_{i=1}^n r_g(t_i) \nabla s_g(t_i)$, où $r_g(t_i)$ est défini dans le modèle (2.5) et $\nabla s_g(t_i)$ est défini en (4.82). Alors

$$\lim_{n \rightarrow \infty} \mu_{\theta_g, n} = 0.$$

Preuve. Dans la partie d'introduction portant sur la théorie de Fourier, nous avons vu que pour une fonction $g(t)$ périodique de période $P_0 = \frac{2\pi}{\omega_0}$,

$$\begin{aligned} g(t) &= \sum_{k=-\infty}^{\infty} c_{g,k} \exp(i\omega_0 t) \\ &= c_{g,0} + c_{g,1} \exp(i\omega_0 t) + c_{g,-1} \exp(-i\omega_0 t) \\ &\quad + \sum_{k=-\infty}^{-2} c_{g,k} \exp(ik\omega_0 t) + \sum_{k=2}^{\infty} c_{g,k} \exp(ik\omega_0 t) \\ &= c_{g,0} + s_g(t) + r_g(t), \end{aligned} \quad (4.83)$$

où

$$r_g(t) = \sum_{k=-\infty}^{-2} c_{g,k} \exp(ik\omega_0 t) + \sum_{k=2}^{\infty} c_{g,k} \exp(ik\omega_0 t),$$

et $c_{g,0} = \mu_g$, $c_{g,1} = A_g - iB_g$, $c_{g,-1} = A_g + iB_g$. Par la P_0 périodicité des fonctions $r_g(t)$ et $s_g(t)$, nous pouvons donc considérer les temps $\tilde{t}_{(i)} = t_i - k_i P_0$, $i = 1, \dots, n$ et $k_i \in \mathbb{N}$, repliés en phase et ordonnés, selon la définition 1. Considérons la première composante de (4.81), $-\sum_{i=1}^n (y_i - s_g(t_i) (B_g \cos(\omega_0 t_i) t_i - A_g \sin(\omega_0 t_i) t_i))$. Avec les temps $\tilde{t}_{(i)}$, on obtient

$$\sum_{i=1}^n r_g(\tilde{t}_{(i)}) (A_g \cos(\omega_0 \tilde{t}_{(i)}) (\tilde{t}_{(i)} + k_i P_0) - B_g \sin(\omega_0 \tilde{t}_{(i)}) (\tilde{t}_{(i)} + k_i P_0)). \quad (4.84)$$

Lorsque $n \rightarrow \infty$, on observe un très grand nombre de fois $\cos(\omega_0 \tilde{t}_{(i)}) = \cos(\omega_0 t_u)$, à différents temps d'observations t_u , $u \in \{1, \dots, n\}$. Il existe donc plusieurs $k \in \mathbb{N}$, tels que $t_u = \tilde{t}_{(i)} + k P_0$, $k \in \mathbb{N}$. De même pour $r_g(t)$. Ainsi (4.84) devient

$$n \sum_{i=1}^n r_g(\tilde{t}_i) \left(A_g \cos(\omega_0 \tilde{t}_i) \frac{\tilde{t}_i + k_i P_0}{n} - B_g \sin(\omega_0 \tilde{t}_i) \frac{\tilde{t}_i + k_i P_0}{n} \right). \quad (4.85)$$

Et (4.85) vaut

$$n \frac{n}{P_0} \int_0^{P_0} r_g(u) (A_g \cos(\omega_0 u) - B_g \sin(\omega_0 u)) du \sum_{k=1}^n \frac{k P_0}{n}. \quad (4.86)$$

Pour $n \rightarrow \infty$, on a que $\frac{k}{n} \rightarrow w$, avec $0 < w < 1$. Et donc $n \sum_{k=0}^n \frac{k}{n} \frac{1}{n} \simeq n \int_0^1 w dw = \frac{n}{2}$. Donc (4.87) devient

$$n^3 \int_0^{P_0} r_g(u) (A_g \cos(\omega_0 u) - B_g \sin(\omega_0 u)) du \int_0^1 w du.$$

Premièrement, considérons une composante complexe $r_g^k(t) + r_g^{-k}(t) = c_k \exp(ik\omega_0 t) + c_{-k} \exp(-ik\omega_0 t)$ de la fonction $r_g(t)$. Alors,

$$\int_0^{P_0} (r_g^k(t) + r_g^{-k}(t)) (A_g \cos(\omega_0 t) - B_g \sin(\omega_0 t)) dt. \quad (4.87)$$

Considérons l'intégrale de (4.87) avec la composante complexe $r_g^k(t)$. Les calculs pour la composante complexe $r_g^{-k}(t)$ sont analogues. Commençons par calculer la première composante $\int_0^{P_0} (r_g^k(t)) (B_g \cos(\omega_0 t) - A_g \sin(\omega_0 t)) dt$ de (4.87). On obtient

$$\int_0^{P_0} \frac{1}{2} \left(c_{g,k} B_g \exp(i(k+1)\omega_0 t) + \frac{1}{2} c_{g,k} B_g \exp(i(k-1)\omega_0 t) \right) dt = 0,$$

et

$$\int_0^{P_0} \frac{1}{2i} \left(c_{g,k} A_g \exp(i(k+1)\omega_0 t) - \frac{1}{2} c_{g,k} A_g \exp(i(k-1)\omega_0 t) \right) dt = 0.$$

En effet, la période de $\exp(i(k+1)\omega_0 t)$ vaut $\frac{P_0}{k+1}$ et donc

$$\int_0^{P_0} \exp(i(k+1)\omega_0 t) dt = \sum_{l=0}^{k+1} \int_{l \frac{P_0}{k+1}}^{(l+1) \frac{P_0}{k+1}} \exp(i(k+1)\omega_0 t) dt,$$

qui est équivalent à

$$\int_0^{\frac{P_0}{k+1}} \exp(i(k+1)\omega_0 t) dt = \int_0^{\frac{P_0}{k+1}} \cos((k+1)\omega_0 t) dt + i \int_0^{\frac{P_0}{k+1}} \sin((k+1)\omega_0 t) dt, \quad (4.88)$$

et l'intégrale des fonctions cosinus et sinus sur leur période vaut zéro. De même pour $\exp(i(k-1)\omega_0 t)$. On obtient les mêmes résultats asymptotiques, pour les deux autres composantes de (4.81). Ainsi, $\lim_{n \rightarrow \infty} \sum_{i=1}^n r_g(t_i) \nabla s_g(t_i) = 0$. ■

Il nous reste à calculer la variance asymptotique $n\Sigma_{\theta_g}$ de $\sqrt{n}(\hat{\theta}_n - \theta_g)$,

$$n\Sigma_{\hat{\theta}_n} \simeq \text{Var} \left(\left(-\frac{1}{n} H(\text{SC}(\theta_g)) \right)^{-1} \frac{1}{\sqrt{n}} \nabla \text{SC}(\theta_g) \right).$$

En utilisant le fait que si $n \rightarrow \infty$, alors $\frac{1}{n} H(\text{SC}(\theta_g)) \rightarrow E(H(\text{SC}(\theta_g)))$. La matrice Hessienne $H(\text{SC}(\theta_g))$ de la somme des carrés vaut

$$\begin{aligned} H(\text{SC}(\theta_g)) &= \nabla \nabla^T \text{SC}(\theta_g) \\ &= -\nabla \sum_{i=1}^n (y_i - s_g(t_i)) \nabla^T s_g(t_i) \\ &= \sum_{i=1}^n \nabla s_g(t_i) \nabla^T s_g(t_i) + \sum_{i=1}^n (y_i - s_g(t_i)) H(s_g(t_i)). \end{aligned} \quad (4.89)$$

Et donc l'espérance $\frac{1}{n} H(\text{SC}(\theta_g))$ tend vers $\mathbb{E}[H(\text{SC}(\theta_g))]$, lorsque $n \rightarrow \infty$,

$$\begin{aligned} \mathbb{E}[H(\text{SC}(\theta_g))] &\simeq \frac{1}{n} H(\text{SC}(\theta_g)) = \frac{1}{n} \sum_{i=1}^n \nabla s_g(t_i) \nabla^T s_g(t_i) + \frac{1}{n} \sum_{i=1}^n (y_i - s_g(t_i)) H(s_g(t_i)) \\ &= \frac{1}{n} \sum_{i=1}^n \nabla s_g(t_i) \nabla^T s_g(t_i) + \frac{1}{n} \sum_{i=1}^n r_g(t_i) H(s_g(t_i)). \end{aligned} \quad (4.90)$$

En utilisant la relation

$$H(\text{SC}(\theta_g)) = \left(\frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_j} \text{SC}(\theta_g) \right)_{i,j=1,\dots,4}.$$

Posons $H^*(\text{SC}(\theta_g)) = \sum_{i=1}^n \nabla s_g(t_i) \nabla^T s_g(t_i)$. On obtient explicitement

$$\begin{aligned} H^*(\text{SC}(\theta)) &= \sum_{i=1}^n \nabla s_g(t_i) \nabla^T s_g(t_i) \\ &= \begin{bmatrix} \sum_{i=1}^n C_i^2 & \sum_{i=1}^n \cos(\omega_0 t_i) C_i & \sum_{i=1}^n \sin(\omega_0 t_i) C_i \\ \sum_{i=1}^n \cos(\omega_0 t_i) C_i & \sum_{i=1}^n \cos(\omega_0 t_i)^2 & \sum_{i=1}^n \cos(\omega_0 t_i) \sin(\omega_0 t_i) \\ \sum_{i=1}^n \sin(\omega_0 t_i) C_i & \sum_{i=1}^n \cos(\omega_0 t_i) \sin(\omega_0 t_i) & \sum_{i=1}^n \sin(\omega_0 t_i)^2 \end{bmatrix}, \end{aligned} \quad (4.91)$$

où $C_i = (B_g \cos(\omega_0 t_i) - A_g \sin(\omega_0 t_i)) t_i$, et

$$\begin{aligned}
R_{\theta_g, n} &= \sum_{i=1}^n r_g(t_i) H(s_g(t_i)) \\
&= \begin{bmatrix} \sum_{i=1}^n r_g(t_i) D_i & -\sum_{i=1}^n r_g(t_i) \sin(\omega_0 t_i) t_i & \sum_{i=1}^n r_g(t_i) \cos(\omega_0 t_i) t_i \\ -\sum_{i=1}^n r_g(t_i) \sin(\omega_0 t_i) t_i & 0 & 0 \\ \sum_{i=1}^n r_g(t_i) \cos(\omega_0 t_i) t_i & 0 & 0 \end{bmatrix}, \quad (4.92)
\end{aligned}$$

où $D_i = -(B_g \cos(\omega_0 t_i) + A_g \sin(\omega_0 t_i)) t_i^2$.

$$\begin{aligned}
n\Sigma_\theta &\simeq E \left(H(\text{SC}(\hat{\theta}))^{-1} \right) \text{Var} \left(\frac{1}{\sqrt{n}} \nabla \text{SC}(\theta) \right) E \left(H(\text{SC}(\hat{\theta}))^{-1} \right)^t \\
&\simeq \frac{1}{n} \left(\frac{1}{n} H^*(\text{SC}(\hat{\theta})) \right)^{-1} \text{Var} (\nabla \text{SC}(\theta)) \left(\frac{1}{n} H^*(\text{SC}(\hat{\theta})) \right)^{-1} \\
&\simeq n H^*(\text{SC}(\hat{\theta}))^{-1} \text{Var} [\nabla \text{SC}(\theta)] H^*(\text{SC}(\hat{\theta}))^{-1}. \quad (4.93)
\end{aligned}$$

A la ligne (4.93), on utilise le fait que la matrice Hessienne est symétrique et que l'espérance et l'inversion d'une matrice sont commutatifs.

Lemme 3 *Soit g une fonction périodique de période P_0 , vérifiant les conditions (4.2), (4.3) et (4.4). L'échantillonnage est irrégulier, mais vérifiant la définition 2 de la couverture de la période. De plus, $\eta = \lim_{i \rightarrow \infty} \mathbb{E}[T_i - T_{i-1}]$ existe. Alors la matrice $R_{\theta_g, n}$, définie au point (4.99) ci-dessus, converge, lorsque $n \rightarrow \infty$, vers la matrice nulle.*

Preuve. Nous considérons à nouveau le développement du signal périodique en série de Fourier, comme dans le lemme 2. Aussi, considérons une composante complexe $r_g^k(t) + r_g^{-k}(t) = c_{g,k} \exp(ik\omega_0 t) + c_{g,-k} \exp(-ik\omega_0 t)$ de la fonction $r_g(t)$ pour la composante $R_{11} = \sum_{i=1}^n r_g(t_i) D_i$ de la matrice $R_{\theta_g, n}$. Comme dans la preuve du lemme 2, considérons les temps repliés en phase et ordonnés $\tilde{t}_{(i)}$, $i = 1, \dots, n$. Remarquons que $D_i = -s_g(t_i) t_i^2$. On peut donc écrire $[R_{\theta_g, n}]_{11}$ comme

$$\begin{aligned}
& - \sum_{i=1}^n \left(r_g^k(\tilde{t}_{(i)}) + r_g^{-k}(\tilde{t}_{(i)}) \right) s_g(\tilde{t}_{(i)}) (\tilde{t}_{(i)} + k_i P_0)^2 \\
& = -n^2 \sum_{i=1}^n \left(r_g^k(\tilde{t}_{(i)}) + r_g^{-k}(\tilde{t}_{(i)}) \right) s_g(\tilde{t}_{(i)}) \left(\frac{\tilde{t}_{(i)} + k_i P_0}{n} \right)^2. \quad (4.94)
\end{aligned}$$

Lorsque $n \rightarrow \infty$, on observe un très grand nombre de fois $\cos(\omega_0 \tilde{t}_{(i)}) = \cos(\omega_0 t_u)$, à différents temps d'observations t_u , $u \in \{1, \dots, n\}$. Il existe donc plusieurs $k \in \mathbb{N}$, tels que $t_u = \tilde{t}_{(i)} + kP_0$, $k \in \mathbb{N}$. De même pour $r_g(t)$. Ainsi (4.94) devient

$$- \frac{n^3}{P_0} \int_0^{P_0} \left(r_g^k(t) + r_g^{-k}(t) \right) (s_g(t)) dt \sum_{k=1}^n \left(\frac{k}{n} \right)^2. \quad (4.95)$$

Pour $n \rightarrow \infty$, on a que $\frac{k}{n} \rightarrow w$, avec $0 < w < 1$. Et donc $n \sum_{k=0}^n \left(\frac{k}{n} \right)^2 \frac{1}{n} \simeq n \int_0^1 w^2 dw = \frac{n}{3}$. Donc (4.95) devient

$$- \frac{1}{3} \frac{n^4}{P_0} \int_0^{P_0} \left(r_g^k(t) + r_g^{-k}(t) \right) (s_g(t)) dt. \quad (4.96)$$

Concentrons nous sur l'intégrale de la composante $r_\theta^k(t)$ de (4.96).

$$\begin{aligned} -\frac{1}{3} \frac{n^4}{P_0} \int_0^{P_0} r_g^k(t) s_g(t) dt &= \frac{1}{3} \frac{n^4}{P_0} \int_0^{P_0} r_g^k(t) (A_g \cos(\omega_0 t) + B_g \sin(\omega_0 t)) dt \\ &= \frac{1}{3} \frac{n^4}{P_0} \int_0^{P_0} r_g^k(t) (A_g \cos(\omega_0 t) + B_g \sin(\omega_0 t)) dt. \end{aligned} \quad (4.97)$$

Considérons l'intégrale $\frac{n^4}{P_0} \int_0^{P_0} r_g^k(t) A_g \cos(\omega_0 t) dt \frac{1}{3}$ de (4.97). En remplaçant $\cos(\omega_0 t) = \frac{\exp(i\omega_0 t) + \exp(-i\omega_0 t)}{2}$, on obtient

$$-\frac{1}{3} \frac{n^4}{P_0} \int_0^{P_0} r_g^k(t) A_g \frac{\exp(i\omega_0 t) + \exp(-i\omega_0 t)}{2} dt. \quad (4.98)$$

Comme dans la preuve du lemme 2, $r_g^k(t) = c_k \exp(ik\omega_0 t)$. (4.98) devient

$$-\frac{1}{3} \frac{n^4}{P_0} \int_0^{P_0} \frac{1}{2i} \left(c_k A_g \exp(i(k+1)\omega_0 t) - \frac{1}{2} c_k A_g \exp(i(k-1)\omega_0 t) \right) dt = 0,$$

pour les mêmes raisons données dans la preuve du lemme 2. On obtient ainsi le même résultat pour toutes les autres composantes $r_g^k(t)$, $k \in \mathbb{Z}$ de la série de Fourier. Ainsi, les fonctions $r_g(t)$ et $g_{\theta_g}(t)$ sont orthogonales sur l'intervalle $[0, P_0]$. En effectuant les mêmes types de calculs, on obtient les mêmes résultats asymptotiques pour les autres composantes de la matrice R. \blacksquare

Par le lemme 3, $E(H(SC(\theta_g)))$ vaut, lorsque $n \rightarrow \infty$,

$$E(H(SC(\theta_g))) = \frac{1}{n} H^*(SC(\theta_g))$$

qui est la matrice Hessienne du modèle de régression sinusoidale avec l'hypothèse que le signal est de forme sinusoidale. Dans la suite de la preuve, nous approximations les sommes par des intégrales de Riemann. Dans les hypothèses du théorème, la distance entre deux temps successifs vaut en moyenne $\eta = 1$.

Lemme 4 Pour $n \rightarrow \infty$, $H^*(SC(\theta_g))$, (4.99), tend vers la matrice

$$\begin{bmatrix} \frac{1}{6} n^3 (A_g^2 + B_g^2) + O(n^2) & n^2 B_g \frac{1}{4} + O(n) & -n^2 A_g \frac{1}{4} + O(n) \\ n^2 B_g \frac{1}{4} + O(n) & \frac{1}{2} n & O(1) \\ -n^2 A_g \frac{1}{4} + O(n) & O(1) & \frac{1}{2} n \end{bmatrix}.$$

Preuve. Effectuer les calculs pour chaque composante.

$$\begin{aligned}
\sum_{i=1}^n C_i^2 &= \int_0^n (-A_g \sin(\omega_0 t) + B_g \cos(\omega_0 t))^2 t^2 dt + O(1) \\
&= A_g^2 \int_0^n \sin(\omega_0 t)^2 t^2 dt - 2A_g B_g \int_0^n \sin(\omega_0 t) \cos(\omega_0 t) t^2 dt \\
&\quad + B_g^2 \int_0^n \cos(\omega_0 t)^2 t^2 dt + O(1) \\
&= A_g^2 \int_0^n \frac{1 - \cos(2\omega_0 t)}{2} t^2 dt - A_g B_g \int_0^n \sin(2\omega_0 t) t^2 dt \\
&\quad + B_g^2 \int_0^n \frac{1 + \cos(2\omega_0 t)}{2} t^2 dt + O(1) \\
&= (A_g^2 + B_g^2) \frac{1}{6} n^3 + B_g^2 \frac{1}{6} n^3 + O(n^2), \\
\sum_{i=1}^n \cos(\omega_0 t_i) C_i &= \int_0^n (A_g \sin(\omega_0 t) + B_g \cos(\omega_0 t)) t \cos(\omega_0 t) dt + O(1) \\
&= -A_g \int_0^n \sin(\omega_0 t) \cos(\omega_0 t) t dt + B_g \int_0^n \cos(\omega_0 t)^2 t dt + O(1) \\
&= B_g \int_0^n \frac{1 + \cos(2\omega_0 t)}{2} t dt + O(n) \\
&= B_g \frac{1}{4} n^2 + O(n), \\
\sum_{i=1}^n \sin(\omega_0 t_i) C_i &= \int_0^n (-A_g \sin(\omega_0 t) + B_g \cos(\omega_0 t)) t \sin(\omega_0 t) dt + O(1) \\
&= -A_g \int_0^n \sin(\omega_0 t)^2 t dt + B_g \int_0^n \cos(\omega_0 t) \sin(\omega_0 t) t dt + O(1) \\
&= -A_g \int_0^n \frac{1 - \cos(2\omega_0 t)}{2} t dt + O(n) \\
&= -A_g \frac{1}{4} n^2 + O(n),
\end{aligned}$$

$$\begin{aligned}
\sum_{i=1}^n \cos(\omega_0 t_i)^2 &= \int_0^n \cos(\omega_0 t)^2 dt \\
&= \frac{n}{2} + O(1), \\
\sum_{i=1}^n \cos(\omega_0 t_i) \sin(\omega_0 t_i) &= \int_0^n \sin(\omega_0 t) \cos(\omega_0 t) dt \\
&= O(1), \\
\sum_{i=1}^n \sin(\omega_0 t_i)^2 &= \int_0^n \sin(\omega_0 t)^2 dt \\
&= \frac{n}{2} + O(1).
\end{aligned}$$

■

Il nous faut encore calculer la matrice $\text{Var}(\nabla \text{SC}(\theta_g))$.

Lemme 5 *Soit g une fonction périodique de période P_0 , vérifiant les conditions (4.2), (4.3) et (4.4). L'échantillonnage est irrégulier, mais vérifiant la définition 2 de la couverture de la période. De plus,*

$\eta = \lim_{i \rightarrow \infty} \mathbb{E}[T_i - T_{i-1}]$ existe. On suppose le processus Gaussien des erreurs stationnaire et ergodique. On suppose aussi que la fonction de densité de spectrale $f_{sp}(\omega)$ existe. Le processus a une fonction d'autocovariance $\gamma(x) \in \mathbb{R}$ non forcément nulle pour $x \in \mathbb{R}$. Alors, lorsque $n \rightarrow \infty$,

$$\text{Var}[\nabla SC(\theta_g)]_{ij} = 2 \int_{-\infty}^{\infty} \mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,j}} s_g \right\}(\omega) \overline{\mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,i}} s_g \right\}(\omega)} f_{sp}(\omega) d\omega,$$

$i, j = 1, \dots, 3$. De plus, si $i = j$,

$$\text{Var}[\nabla SC(\theta)]_{ii} = 2 \int_{-\infty}^{\infty} \left| \mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,i}} s_g \right\}(\omega) \right|^2 f_{sp}(\omega) d\omega.$$

Preuve.

$$\begin{aligned} \text{Var}[\nabla SC(\theta_g)] &= \text{Cov}[\nabla SC(\theta_g), \nabla SC(\theta_g)] \\ &= \text{Cov} \left[\sum_{i=1}^n (Y_i - s_g(t_i)) \nabla s_g(t_i), \sum_{j=1}^n (Y_j - s_g(t_j)) \nabla s_g(t_j) \right] \\ &= \sum_{i=1}^n \sum_{j=1}^n \nabla s_g(t_i) \nabla^T s_g(t_j) \gamma(|t_i - t_j|). \end{aligned} \tag{4.99}$$

Quand $n \rightarrow \infty$, on a pour la composante ij de la variance

$$\begin{aligned} [\text{Var}_{asy}(\nabla SC(\theta_g))]_{ij} &= 2 \int_0^n \int_0^{n-x} H_{ij}(x, x+y) \gamma(y) dy dx + O(n^{-1}) \\ &= 2 \int_0^n \frac{\partial}{\partial \theta_{g,i}} s_g(x) \int_0^{n-x} \frac{\partial}{\partial \theta_{g,j}} s_g(x+y) \gamma(y) dy dx + O(n^{-1}) \\ &= 2 \int_0^n \frac{\partial}{\partial \theta_{g,i}} s_g(x) \frac{\partial}{\partial \theta_{g,j}} s_g \star \gamma(x) dx + O(n^{-1}) \\ &= 2 \int_0^n \frac{\partial}{\partial \theta_{g,i}} s_g(x) \int_{-\infty}^{\infty} \mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,j}} s_g \right\}(\omega) f_{sp}(\omega) \exp(i\omega x) d\omega dx + O(n^{-1}) \\ &= 2 \int_{-\infty}^{\infty} \mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,j}} s_g \right\}(\omega) f_{sp}(\omega) \int_0^n \frac{\partial}{\partial \theta_{g,i}} s_g(x) \exp(i\omega x) dx d\omega + O(n^{-1}) \\ &= 2 \int_{-\infty}^{\infty} \mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,j}} s_g \right\}(\omega) f_{sp}(\omega) \overline{\mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,i}} s_g \right\}(\omega)} d\omega + O(n^{-1}) \\ &= 2 \int_{-\infty}^{\infty} \mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,j}} s_g \right\}(\omega) \overline{\mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,i}} s_g \right\}(\omega)} f_{sp}(\omega) d\omega + O(n^{-1}), \end{aligned} \tag{4.100}$$

avec $f_{sp}(\omega) = \int_{-\infty}^{+\infty} \gamma(s) \exp(-i\omega s) ds$ la densité spectrale (1.5) définie au chapitre 1.5.

$$H_{ij}(x, x+y) = [\nabla s_g(x) \nabla^T s_g(x+y)]_{ij} = \frac{\partial}{\partial \theta_{g,i}} s_g(x) \frac{\partial}{\partial \theta_{g,j}} s_g(x+y),$$

avec $x, y \geq 0$. Si $i = j$, on a

$$[\text{Var}_{asy}(\nabla \text{SC}(\theta_g))]_{ii} \simeq 2 \int_{-\infty}^{\infty} \left| \mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,i}} s_g \right\}(\omega) \right|^2 f_{sp}(\omega) d\omega.$$

■

Appliquons le résultat du lemme 5, à la fonction de régression sinusoidale. Commençons par calculer les transformées de Fourier suivantes.

$$\begin{aligned} \mathcal{TF} \{ \cos(\omega_0 t) \} &= \frac{1}{2} (\delta(\omega - \omega_0) + \delta(\omega + \omega_0)), \\ \mathcal{TF} \{ \sin(\omega_0 t) \} &= \frac{1}{2i} (\delta(\omega - \omega_0) - \delta(\omega + \omega_0)), \\ \mathcal{TF} \{ \cos(\omega_0 t) t \} &= \frac{1}{2} \int_{-\infty}^{\infty} t \exp(-i(\omega - \omega_0)t) dt + \frac{1}{2} \int_{-\infty}^{\infty} t \exp(-i(\omega + \omega_0)t) dt \\ &= \frac{1}{-2i} \frac{\partial}{\partial \omega} \int_{-\infty}^{\infty} \exp(-i(\omega - \omega_0)t) dt + \frac{1}{-2i} \frac{\partial}{\partial \omega} \int_{-\infty}^{\infty} \exp(-i(\omega + \omega_0)t) dt \\ &= \frac{1}{-2i} \frac{\partial}{\partial \omega} (\delta(\omega - \omega_0) + \delta(\omega + \omega_0)) \\ &= \frac{1}{-2i} (\delta'(\omega - \omega_0) + \delta'(\omega + \omega_0)). \end{aligned}$$

De la même façon, on obtient que

$$\mathcal{TF} \{ \sin(\omega_0 t) t \} = \frac{1}{2} (\delta'(\omega - \omega_0) - \delta'(\omega + \omega_0)),$$

où

$$\delta(\omega) = \begin{cases} +\infty & \text{si } \omega = 0 \\ 0 & \text{si } \omega \neq 0 \end{cases},$$

est la fonction de Dirac.

$$\mathcal{TF} \left(\frac{\partial}{\partial \theta_{g,1}} s_g \right) (\omega) = B_{g,1} \mathcal{TF} \{ \cos(\omega_0 t) t \} - A_{g,1} \mathcal{TF} \{ \sin(\omega_0 t) t \}.$$

Et donc

$$\begin{aligned} \left| \mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,1}} s_g \right\}(\omega) \right|^2 &= \frac{A_{g,1}^2}{4} (\delta'(\omega - \omega_0) + \delta'(\omega + \omega_0))^2 \\ &\quad - \frac{A_{g,1} B_{g,1}}{2} (\delta'(\omega - \omega_0) + \delta'(\omega + \omega_0)) (\delta'(\omega - \omega_0) - \delta'(\omega + \omega_0)) \\ &\quad + \frac{B_{g,1}^2}{4} (\delta'(\omega - \omega_0) - \delta'(\omega + \omega_0))^2. \end{aligned}$$

L'intégrale donne

$$\begin{aligned} \int_{-\infty}^{\infty} \left| \mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,1}} s_g \right\}(\omega) \right|^2 f_{sp}(\omega) d\omega &= \frac{B_{g,1}^2}{4} \int_{-\infty}^{\infty} (\delta'(\omega - \omega_0) + \delta'(\omega + \omega_0))^2 f_{sp}(\omega) d\omega \\ &\quad - \frac{A_{g,1} B_{g,1}}{2} \int_{-\infty}^{\infty} (\delta'(\omega - \omega_0) + \delta'(\omega + \omega_0)) (\delta'(\omega - \omega_0) - \delta'(\omega + \omega_0)) f_{sp}(\omega) d\omega \\ &\quad + \frac{A_{g,1}^2}{4} \int_{-\infty}^{\infty} (\delta'(\omega - \omega_0) - \delta'(\omega + \omega_0))^2 f_{sp}(\omega) d\omega. \end{aligned}$$

La première intégrale s'approxime, pour $n \rightarrow \infty$, par

$$-\frac{B_{g,1}^2}{2i} \int_{-\infty}^{\infty} \left(\delta'(\omega - \omega_0) + \delta'(\omega + \omega_0) \right) \overline{\int_0^n \cos(\omega_0 t) t \exp(-i\omega t) dt} f_{sp}(\omega) d\omega,$$

ce qui devient

$$-\frac{B_{g,1}^2}{2i} \int_0^n \cos(\omega_0 t) t \int_{-\infty}^{\infty} \left(\delta'(\omega - \omega_0) + \delta'(\omega + \omega_0) \right) \exp(i\omega t) f_{sp}(\omega) d\omega dt.$$

Et l'intégrale sur la variable ω donne

$$\begin{aligned} \int_{-\infty}^{\infty} \left(\delta'(\omega - \omega_0) + \delta'(\omega + \omega_0) \right) \exp(i\omega t) f_{sp}(\omega) d\omega \\ &= -\frac{\partial}{\partial \omega} (\exp(i\omega t) f_{sp}(\omega))_{\omega=\omega_0} - \frac{\partial}{\partial \omega} (\exp(i\omega t) f_{sp}(\omega))_{\omega=-\omega_0} \\ &= -\left(it \exp(i\omega_0 t) f_{sp}(\omega_0) + f'_{sp}(\omega_0) \exp(i\omega_0 t) \right) \\ &\quad - \left(it \exp(-i\omega_0 t) f_{sp}(\omega_0) + f'_{sp}(\omega_0) \exp(-i\omega_0 t) \right) \\ &= -f_{sp}(\omega_0) it (\exp(i\omega_0 t) + \exp(-i\omega_0 t)) \\ &\quad - f'_{sp}(\omega_0) (\exp(i\omega_0 t) + \exp(-i\omega_0 t)) \\ &= -f_{sp}(\omega_0) it 2 \cos(\omega_0 t) - f'_{sp}(\omega_0) 2 \cos(\omega_0 t) \\ &= -2 \cos(\omega_0 t) \left(f_{sp}(\omega_0) it + f'_{sp}(\omega_0) \right). \end{aligned}$$

Ainsi l'intégrale sur la variable de temps t devient

$$\begin{aligned} \frac{B_g^2}{2i} \int_0^n \cos(\omega_0 t) t 2 \cos(\omega_0 t) \left(f_{sp}(\omega_0) it + f'_{sp}(\omega_0) \right) dt \\ &= B_g^2 f_{sp}(\omega_0) \int_0^n \cos(\omega_0 t)^2 t^2 dt - A_g^2 f'_{sp}(\omega_0) \frac{1}{i} \int_0^n \cos(\omega_0 t)^2 t dt \\ &= B_g^2 f_{sp}(\omega_0) \int_0^n \frac{1 + \cos(2\omega_0 t)}{2} t^2 dt \\ &\quad - A_g^2 f'_{sp}(\omega_0) i \int_0^n \frac{1 + \cos(2\omega_0 t)}{2} t dt \\ &= B_g^2 f_{sp}(\omega_0) \frac{1}{6} n^3 + O(n^2), \end{aligned}$$

où on utilise une intégrale par partie pour trouver que

$$\begin{aligned} \int_0^n \frac{\cos(2\omega_0 t)}{2} t^2 dt &= O(n^2) \\ \int_0^n \frac{1 + \cos(2\omega_0 t)}{2} t dt &= O(n^2). \end{aligned}$$

Avec les mêmes type de calculs, et en utilisant que,

$$\begin{aligned}
& \int_{-\infty}^{\infty} \left(\delta'(\omega - \omega_0) - \delta'(\omega + \omega_0) \right) \exp(i\omega t) f_{sp}(\omega) d\omega \\
&= -\frac{\partial}{\partial \omega} (\exp(i\omega t) f_{sp}(\omega))_{\omega=\omega_0} + \frac{\partial}{\partial \omega} (\exp(i\omega t) f_{sp}(\omega))_{\omega=-\omega_0} \\
&= -\left(it \exp(i\omega_0 t) f_{sp}(\omega_0) + f'_{sp}(\omega_0) \exp(i\omega_0 t) \right) \\
&\quad + \left(it \exp(-i\omega_0 t) f_{sp}(\omega_0) + f'_{sp}(\omega_0) \exp(-i\omega_0 t) \right) \\
&= -f_{sp}(\omega_0) it (\exp(i\omega_0 t) - \exp(-i\omega_0 t)) \\
&\quad - f'_{sp}(\omega_0) (\exp(i\omega_0 t) - \exp(-i\omega_0 t)) \\
&= -f_{sp}(\omega_0) it 2i \sin(\omega_0 t) - f'_{sp}(\omega_0) 2i \sin(\omega_0 t) \\
&= -2i \sin(\omega_0 t) \left(f(\omega_0) it + f'_{sp}(\omega_0) \right),
\end{aligned}$$

on obtient

$$\frac{A_g^2}{4} \int_{-\infty}^{\infty} \left(\delta'(\omega - \omega_0) - \delta'(\omega + \omega_0) \right)^2 f_{sp}(\omega) d\omega \simeq A_g^2 f_{sp}(\omega) \frac{1}{6} n^3 + O(n^2). \quad (4.101)$$

On approxime l'intégrale correspondant aux termes mixtes AB avec $n \rightarrow \infty$, par

$$\begin{aligned}
-\frac{A_g B_g}{2} \int_0^n \sin(\omega_0 t) it 2 \cos(\omega_0 t) \left(f_{sp}(\omega_0) it + f'_{sp}(\omega_0) \right) dt &= A_g B_g f_{sp}(\omega_0) \int_0^n \frac{\sin(2\omega_0 t)}{2} t^2 dt \\
&\quad - A_g B_g f'_{sp}(\omega_0) i \int_0^n \frac{\sin(2\omega_0 t)}{2} t dt \\
&= O(n^2).
\end{aligned}$$

Le résultat s'obtient à nouveau en effectuant une intégrale par partie. Finalement on obtient

$$\int_{-\infty}^{\infty} \left| \mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,1}} s_g \right\} (\omega) \right|^2 f_{sp}(\omega) d\omega = \frac{1}{6} (A_g^2 + B_g^2) f_{sp}(\omega_0) n^3 + O(n^2). \quad (4.102)$$

On fait de même pour les autres composantes de la matrice et on obtient

$$\begin{aligned}
\int_{-\infty}^{\infty} \mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,1}} s_g \right\} (\omega) \overline{\mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,2}} s_g \right\} (\omega)} f_{sp}(\omega) d\omega &= B_g f_{sp}(\omega_0) \frac{1}{4} n^2 + O(n), \\
\int_{-\infty}^{\infty} \mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,1}} s_g \right\} (\omega) \overline{\mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,3}} s_g \right\} (\omega)} f_{sp}(\omega) d\omega &= -A_g f_{sp}(\omega_0) \frac{1}{4} n^2 + O(n), \\
\int_{-\infty}^{\infty} \mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,2}} s_g \right\} (\omega) \overline{\mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,1}} s_g \right\} (\omega)} f_{sp}(\omega) d\omega &= B_g f_{sp}(\omega_0) \frac{1}{4} n^2 + O(n), \\
\int_{-\infty}^{\infty} \mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,2}} s_g \right\} (\omega) \overline{\mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,2}} s_g \right\} (\omega)} f_{sp}(\omega) d\omega &= f_{sp}(\omega_0) \frac{n}{2} + O(1), \\
\int_{-\infty}^{\infty} \mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,2}} s_g \right\} (\omega) \overline{\mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,3}} s_g \right\} (\omega)} f_{sp}(\omega) d\omega &= O(1), \\
\int_{-\infty}^{\infty} \mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,3}} s_g \right\} (\omega) \overline{\mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,1}} s_g \right\} (\omega)} f_{sp}(\omega) d\omega &= -A_g f_{sp}(\omega_0) \frac{1}{4} n^2 + O(n), \\
\int_{-\infty}^{\infty} \mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,3}} s_g \right\} (\omega) \overline{\mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,2}} s_g \right\} (\omega)} f_{sp}(\omega) d\omega &= O(1), \\
\int_{-\infty}^{\infty} \mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,3}} s_g \right\} (\omega) \overline{\mathcal{TF} \left\{ \frac{\partial}{\partial \theta_{g,3}} s_g \right\} (\omega)} f_{sp}(\omega) d\omega &= f_{sp}(\omega_0) \frac{n}{2} + O(1).
\end{aligned}$$

Nous obtenons donc que $\text{Var}_{asy}(\nabla \text{SC}(\theta_g))$ vaut

$$\frac{1}{2\pi} f(\omega_0) \begin{bmatrix} (A_g^2 + B_g^2) \frac{1}{6} n^3 + O(n^2) & B_g \frac{1}{4} n^2 + O(n) & -A_g \frac{1}{4} n^2 + O(n) \\ B_g \frac{1}{4} n^2 + O(n) & \frac{n}{2} + O(1) & O(1) \\ -A_g \frac{1}{4} n^2 + O(n) & O(1) & \frac{n}{2} + O(1) \end{bmatrix}. \quad (4.103)$$

Par le résultat du lemme 4, on peut remarquer que, pour $n \rightarrow \infty$,

$$\text{Var}_{asy}[\nabla \text{SC}(\theta_g)] = \mathbb{E} \left[H(\text{SC}(\hat{\theta}_g)) \right] f_{sp}(\omega_0). \quad (4.104)$$

Par les lemmes 2, 3, 4 et 5, ainsi que les points (4.93), (4.103) et (4.104), nous obtenons que, lorsque $n \rightarrow \infty$, $\left[n^{\frac{3}{2}} (\hat{\omega}_n^{\text{AH}} - \omega_0), n^{\frac{1}{2}} (\hat{A}_n - A_g), n^{\frac{1}{2}} (\hat{B}_n - B_g) \right]^T$ converge en distribution vers une loi Normale multivariée $\mathcal{N}(0, f_{sp}(\omega_0) W_{asy}^{-1})$, avec une matrice de covariances

$$W_{asy} = \begin{bmatrix} \frac{1}{6} (A_g^2 + B_g^2) & B_g \frac{1}{2} & -A_g \frac{1}{2} \\ B_g \frac{1}{2} & \frac{1}{2} & 0 \\ -A_g \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}. \quad (4.105)$$

La variance asymptotique $\text{Var}_{asy} \left[n^{\frac{3}{2}} (\hat{\omega}_n^{\text{AH}} - \omega_0) \right]$ de $n^{\frac{3}{2}} (\hat{\omega}_n^{\text{AH}} - \omega_0)$ vaut

$$f_{sp}(\omega_0) \left((A_g^2 + B_g^2) \frac{1}{6} - W_{asy}(1, (2, 3, 4)) W_{asy}((2, 3, 4), (2, 3, 4))^{-1} W_{asy}((2, 3, 4), 1) \right)^{-1},$$

où

$$\begin{aligned}
W_{asy}(1, (2, 3, 4)) W_{asy}((2, 3, 4), (2, 3, 4))^{-1} W_{asy}((2, 3, 4), 1) &= W_{asy}(1, (2, 3, 4)) \frac{1}{4} (B_{g,1}, -A_{g,1}, 0)^T \\
&= (A_g^2 + B_g^2) \frac{1}{8}.
\end{aligned}$$

Ainsi

$$\begin{aligned}\text{Var}_{asy} \left[n^{\frac{3}{2}} (\widehat{\omega}_n^{\text{AH}} - \omega_0) \right] &= f_{sp}(\omega_0) \left((A_g^2 + B_g^2) \left(\frac{1}{6} - \frac{1}{8} \right) \right)^{-1} \\ &= \frac{24}{A_g^2 + B_g^2} f_{sp}(\omega_0).\end{aligned}\tag{4.106}$$

Dans le cas d'erreurs indépendantes, $f_{sp}(\omega_0) = \sigma^2$, et donc

$$\text{Var}_{asy} \left[n^{\frac{3}{2}} (\widehat{\omega}_n^{\text{AH}} - \omega_0) \right] = \frac{24\sigma^2}{A_g^2 + B_g^2}.$$

■

Influence de la fréquence d'échantillonnage $\frac{1}{\eta}$ sur la variance asymptotique

Le théorème considère des échantillonnages irréguliers avec $\eta = 1$. On peut facilement généraliser les formules avec un $\eta \neq 1$ en effectuant les changements suivants. Considérons les temps aléatoires T_i , $i = 1, \dots, n$ tels que $\mathbb{E}[T_{i+1} - T_i] = 1$. Alors, les temps aléatoires $T'_i = \eta T_i$ vérifient $\mathbb{E}[T'_{i+1} - T'_i] = \eta$. Il suffit de considérer les temps d'observations $\eta t_1, \dots, \eta t_n$ pour obtenir la formule un facteur $\frac{1}{\eta^2}$ dans l'expression de la variance

$$\text{Var}_{asy} \left[n^{\frac{3}{2}} \widehat{\omega}_n^{\text{AH}} \right] = \frac{24}{A_g^2 + B_g^2} f_{sp}(\omega_0) \frac{1}{\eta^2}.$$

4.2.3 Distribution asymptotique de l'estimateur $\widehat{\omega}_n^{\text{NP}}$

Dans cette section, l'échantillonnage est irrégulier, avec la propriété de la couverture de l'intervalle de la période (voir chapitre 1.10 sur l'échantillonnage). Nous considérons une régression non paramétrique sur des observations avec un signal périodique g de période P_0 comme signal sous-jacent, vérifiant les conditions (4.2) et (4.3). Considérons le modèle (2.1), défini au chapitre 2 pour la méthode NP.

$$Y_i = g(t_i) + \varepsilon(t_i),\tag{4.107}$$

où $\varepsilon(t)$ est un processus stochastique en temps continu. L'estimateur $\widehat{\theta} = (\widehat{\omega}_n^{\text{NP}}, \widehat{g}_n)^T$ de $\theta_g = (\omega_0, g)^T$ minimise la somme des carrés (4.130)

$$\text{SC}(\theta = [\omega, g]) = \sum_{i=1}^n (y_i - g(t_i))^2,\tag{4.108}$$

où l'estimateur du signal \widehat{g}_n est la fonction de régression. L'estimateur $\widehat{\omega}_n$ est à calculer à partir des données en effectuant un repliement en phase des données pour chaque fréquence ω à considérer (voir paragraphe (2.4) sur le repliement en phase). Comme nous l'avons vu au paragraphe (2.4), La somme des carrés des résidus présente des minimums locaux aux fractions de la fréquence ω_0 . C'est pourquoi nous considérons que l'estimateur $\widehat{\omega}_n^{\text{NP}}$ se trouve au voisinage de ω_0 . Avant de nous intéresser à la distribution de $\widehat{\omega}_n^{\text{NP}}$ lors d'erreurs corrélées, nous allons exhiber une famille de fonctions continues et périodiques maximisant l'intégrale $\int_0^{P_0} g'(u)^2 du$. L'estimation $\widehat{\omega}_n^{\text{NP}}$ de la fréquence ω_0 sera asymptotiquement meilleure pour un signal périodique $g(t)$ avec une première dérivée $g'(t)$ élevée, ce qui implique de forte croissance et/ou décroissance de la fonction. Un signal avec une faible première dérivée, c'est-à-dire un signal ayant peu d'oscillation, aura une variance asymptotique plus élevée. Il est possible de trouver la fonction qui maximise l'intégrale $\int_0^{P_0} g'(u)^2 du$ sous contrainte que $\int_0^{P_0} g(u)^2 du = C$, avec $C > 0$ une constante.

Théorème 10 *Soit*

$$g(t) = \left(\exp(\sqrt{\lambda t}) + at + b \right), \quad (4.109)$$

où $\lambda > 0$, $a, b \in \mathbb{R}$. Cette fonction maximise l'intégrale $\int_0^{P_0} g'(u)^2 du$ sous contrainte $\int_0^{P_0} g(u)^2 du = C$.

Preuve. Posons l'intégrale à maximiser avec sa contrainte.

$$\max_g \int_0^{P_0} g'(u)^2 du \text{ s.c. } \int_0^{P_0} g(u)^2 du = C,$$

où $C > 0$ est une constante. En utilisant la condition d'Euler-Lagrange pour l'extrema d'une intégrale, qui est

$$\frac{d}{dt} \frac{\partial}{\partial g'} L - \frac{\partial}{\partial g} L = 0,$$

où $L = \int_0^{P_0} g'(u)^2 du + \lambda \int_0^{P_0} g(u)^2 du$ est la fonction de Lagrange, avec multiplicateur de Lagrange λ . La condition d'Euler-Lagrange est

$$2g''(t) - 2\lambda g(t) = 0.$$

Ce qui implique

$$g''(t) = \lambda g(t).$$

Ainsi

$$g(t) = \exp(\sqrt{\lambda t}) + at + b.$$

Et λ est tel que

$$\int_0^{P_0} g(u)^2 du = C.$$

■

Quelques exemples à la figure 4.1 illustrent la fonction (4.109). Le résultat du théorème 11 est argumenté de manière informel, en considérant la procédure du calcul variationnel, comme utilisé dans la preuve du théorème 10.

Théorème 11 (Distribution asymptotique de l'estimateur $\hat{\omega}_n^{\text{NP}}$) *Soit le modèle (4.14), avec erreur $\varepsilon(t)$, un processus stochastique Gaussien stationnaire et ergodique, avec fonction d'autocovariance $\gamma(s)$ vérifiant (4.5). Supposons que la densité spectrale $f_{sp}(\omega)$ de l'erreur existe. La fonction $g(t)$ est périodique de période P_0 vérifiant les conditions (4.2) et (4.3). Soit $\hat{\omega}_n^{\text{NP}}$, l'estimateur de la fréquence ω_0 , défini à la section 2.1. L'échantillonnage est irrégulier, mais vérifiant la définition 2 de la couverture de la période. De plus, $\eta = \lim_{i \rightarrow \infty} \mathbb{E}[T_i - T_{i-1}]$ existe. Alors, lorsque $n \rightarrow \infty$, $n^{\frac{3}{2}} (\hat{\omega}_n^{\text{NP}} - \omega_0)$ converge en distribution vers une loi Normale $\mathcal{N}(0, \tau^2)$, où*

$$\tau^2 = 12\omega_0 \frac{1}{\eta^2} \frac{\int_{-\infty}^{\infty} |\mathcal{TF}\{g'\}(\omega)|^2 f_{sp}(\omega) d\omega}{\left(\int_0^{P_0} g'(u)^2 du \right)^2} \quad (4.110)$$

avec $f_{sp}(\omega)$ la densité spectrale du processus Gaussien corrélé et $\mathcal{TF}\{g'\}$ la transformée de Fourier de la première dérivée du signal.

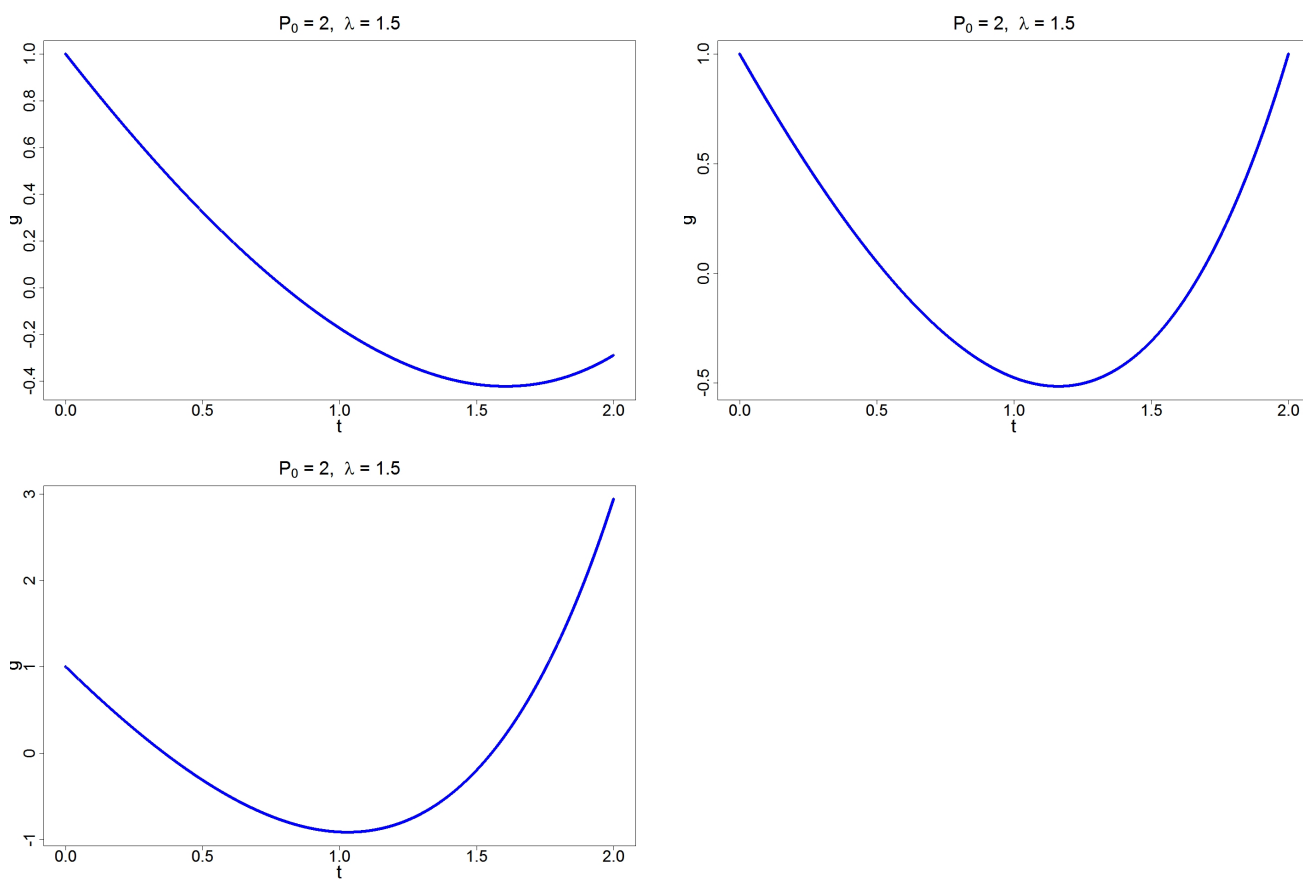


FIG. 4.1 – Exemple de la fonction (4.109) avec période $P_0 = 2$, $a = \frac{2}{P_0^2 \sqrt{\lambda}} \left(1 - \exp\left(\sqrt{\lambda} P_0\right) \right)$, $b = 0$ (fonction centrée) et $\lambda = 0.5, 1, 1.5$ correspondants aux valeurs de $C \simeq 0.37, 0.41, 1.66$, respectivement.

Preuve.

Cette preuve se base sur la preuve donnée dans Hall et al. [13]. Nous donnons un résumé des différentes étapes de la preuve de [13]. Tout comme les auteurs de [13], considérons

$$\widehat{g}(t; P) = \frac{\sum_{j=1}^n Y_j K_j(t; P)}{\sum_{j=1}^n K_j(t; P)} \quad t \in [0, P_0] \quad (4.111)$$

l'estimateur de Nadaraya Watson du signal g , où $K_j(t; P) = \frac{1}{h} K\left(\frac{t - \tilde{t}_j(P)}{h}\right)$ est un noyau et h est la taille du lisseur. Supposons que

1. K est symétrique, à support compact avec trois dérivées continues sur $(-\infty, \infty)$,
2. $h = h(n)$ satisfait $h \sim Cn^{-a}$, où $C > 0$ et $\frac{1}{4} < a < \frac{1}{3}$,
3. g est périodique de période P_0 et vérifie les conditions (4.2) et (4.3),
4. l'échantillonnage vérifie la définition 2 de la couverture de la période. De plus, $\eta = \lim_{j \rightarrow \infty} \mathbb{E}[T_j - T_{j-1}]$.

Séparons le signal $g(t_j)$ des erreurs ε_j en écrivant

$$\widehat{g}(t; P) = \frac{\sum_{j=1}^n g_j K_j(t; P)}{\sum_{j=1}^n n K_j(t; P)} + \frac{\sum_{j=1}^n \varepsilon_j K_j(t; P)}{\sum_{j=1}^n K_j(t; P)}. \quad (4.112)$$

Supposant que $|P - P_0| < n^{-1-\xi}h$, où $\xi > 0$ est fixé. Alors Hall et al. ont montré que si $t \in (2h, P_0 - 2h)$, alors $\tilde{t}_j(P)$ et $\tilde{t}_j(P_0)$ sont tels qu'il existe $k_j \in \mathbb{N}$ vérifiant $\tilde{t}_j(P) = t_j + k_j P$ et $\tilde{t}_j(P_0) = t_j + k_j P_0$, lorsque n est suffisamment grand. Ainsi $\tilde{t}_j(P) - \tilde{t}_j(P_0) = (P - P_0)k_j$. De plus, supposant que K est nulle en dehors de $[-1, 1]$. La première étape est d'effectuer un développement de Taylor de degré 3 de $\widehat{g}(t; P)$ autour de $P = P_0$. Pour cela nous effectuons le développement de Taylor de K_j autour de $P = P_0$.

$$K_j(t; P) = K_j(t; P_0) + K'_j(t; P_0) \frac{k_j}{h} (P - P_0) + \frac{1}{2} K''_j(t; P_0) \frac{k_j^2}{h^2} (P - P_0)^2 + \frac{1}{6} K'''_j(t; P_*) \frac{k_j^3}{h^3} (P - P_0)^3,$$

pour un certain P_* se situant entre P et P_0 et k_j tel que $\tilde{t}_j = t_j + k_j P_0$. Posant $P_n = n \frac{P - P_0}{h}$ et $k_{nj} = \frac{k_j}{n}$. Nous obtenons donc que

$$\begin{aligned} & \widehat{g}(t; P) - \widehat{g}(t; P_0) \\ &= P_n \left[\frac{\sum_{j=1}^n k_{nj} g_j K'_j(t; P_0) \frac{k_j}{h} (P - P_0)}{\sum_{j=1}^n K_j(t; P_0)} - \frac{\sum_{j=1}^n g_j K_j(t; P) \sum_{j=1}^n k_{nj} K_j(t; P)}{\sum_{j=1}^n K_j(t; P) \sum_{j=1}^n K_j(t; P)} \right] \quad (4.113) \end{aligned}$$

$$\begin{aligned} & + O\left(P_n^2 \left[h^2 + (nh)^{-\frac{1}{2}} n^{\xi_1}\right] + |P_n|^3\right) \\ & + P_n \left[\frac{\sum_{j=1}^n k_{nj} \varepsilon_j K'_j(t; P_0) \frac{k_j}{h} (P - P_0)}{\sum_{j=1}^n K_j(t; P_0)} - \frac{\sum_{j=1}^n \varepsilon_j K_j(t; P) \sum_{j=1}^n k_{nj} K_j(t; P)}{\sum_{j=1}^n K_j(t; P) \sum_{j=1}^n K_j(t; P)} \right] \quad (4.114) \\ & + O\left(P_n^2 \left[(nh)^{-\frac{1}{2}} n^{\xi_1}\right] + |P_n|^3 n^{\xi_1}\right). \end{aligned}$$

Lorsque $n \rightarrow \infty$, les auteurs montrent que le terme de la ligne (4.113) multipliant P_n tend vers

$$\frac{1}{2} \frac{1}{P_0} \frac{1}{h} \int_0^{P_0} g(u) K'\left(\frac{t-u}{h}\right) du + O\left(h^2 + (nh)^{-\frac{1}{2}} n^{\xi_1}\right) = \frac{1}{2} \frac{1}{P_0} g'(t) + O\left(h^2 + (nh)^{-\frac{1}{2}} n^{\xi_1}\right), \quad (4.115)$$

pour tout $t \in [2h, P_0 - 2h]$. En effet, lorsque $n \rightarrow \infty$, $h = h(n) \rightarrow 0$. En considérant le changement de variable $v = \frac{t-u}{h}$, cela permet d'effectuer un développement de Taylor de g autour de t , $g(t - vh) =$

$g(t) + g'(t)vh + g''(t)(vh)^2 + O((vh)^2)$. Le terme de la ligne (4.114) multipliant P_n est de l'ordre $O((nh)^{-\frac{1}{2}}n^{\xi_1})$. De plus, $\widehat{g}(t; P_0) = g(t) + \frac{\sum_{j=1}^n \varepsilon_j K_j(t; P_0)}{\sum_{j=1}^n K_j(t; P_0)} + O(h^2 + h(nh)^{-\frac{1}{2}}n^{\xi_1})$. Ainsi,

$$\widehat{g}(t; P) = g(t) + \frac{\sum_{j=1}^n \varepsilon_j K_j(t; P_0)}{\sum_{j=1}^n K_j(t; P_0)} + \frac{1}{2P_0}n(P - P_0)g'(t) + \alpha_1(t; P), \quad (4.116)$$

où $\sup_{t \in [2h, P_0 - 2h]} |\alpha_1(t; P)| = O(\zeta_n(P))$, et $\zeta_n(P) = |P_n|(h^2 + (nh)^{-\frac{1}{2}}n^{\xi_1}) + |P_n|^3 n^{\xi_1} + h^2 + h(nh)^{-\frac{1}{2}}n^{\xi_1}$. La deuxième étape est d'effectuer un développement de Taylor de degré 1 autour de $\tilde{t}(P_0)$. On a que

$$\begin{aligned} \widehat{g}(\tilde{t}(P); P) &= g(\tilde{t}_j(P_0)) + \frac{\sum_{j=1}^n \varepsilon_j K_j(\tilde{t}_j(P_0); P_0)}{\sum_{j=1}^n K_j(\tilde{t}_j(P_0); P_0)} \\ &\quad + n(P - P_0)\left(\frac{1}{2P_0} - \frac{k_j}{n}\right)g'(\tilde{t}_j(P_0)) + \alpha_2(\tilde{t}_j(P_0); P_0), \end{aligned} \quad (4.117)$$

où $\alpha_2(\tilde{t}_j(P_0); P_0)$ satisfait la même condition que $\alpha_1(\tilde{t}_j(P_0); P_0)$. Ainsi, par (4.117),

$$Y_j - \widehat{g}(\tilde{t}_j(P); P) = \varepsilon_j - \frac{\sum_{j=1}^n \varepsilon_j K_j(\tilde{t}_j(P_0); P_0)}{\sum_{j=1}^n K_j(\tilde{t}_j(P_0); P_0)} + n(P - P_0)\left(\frac{k_j}{n} - \frac{1}{2P_0}\right)g'(\tilde{t}_j(P_0)) + \beta_j(P), \quad (4.118)$$

avec $\beta_j(P) = O(|\zeta_n(P)|)$. Par (4.118), on obtient

$$\begin{aligned} SC(P) &= \sum_{j=1}^n (Y_j - \widehat{g}(\tilde{t}_j(P); P))^2 \\ &= \sum_{j=1}^n \left(\varepsilon_j - \frac{\sum_{j=1}^n \varepsilon_j K_j(\tilde{t}_j(P_0); P_0)}{\sum_{j=1}^n K_j(\tilde{t}_j(P_0); P_0)} + n(P - P_0)\left(\frac{k_j}{n} - \frac{1}{2P_0}\right)g'(\tilde{t}_j(P_0)) + \beta_j(P) \right)^2. \end{aligned} \quad (4.119)$$

En dérivant (4.119) par rapport à P pour chercher le minimum de (4.119), on obtient la relation

$$n(P - P_0) = \frac{\sum_{j=1}^n (\frac{k_j}{n} - \frac{1}{2})g'(\tilde{t}_j)\varepsilon_j}{\sum_{j=1}^n (\frac{k_j}{n} - \frac{1}{2})^2 g'(\tilde{t}_j)^2} + o_p(n^{-\frac{1}{2}}). \quad (4.120)$$

Commençons par montrer que, lorsque $n \rightarrow \infty$,

$$\sum_{j=1}^n \left(\frac{k_j}{n} - \frac{1}{2}\right)^2 g'(\tilde{t}_j)^2 = \frac{n}{12} \frac{\eta^2}{P_0^3} \int_0^{P_0} g'(u)^2 du + O(n^{-1}). \quad (4.121)$$

En effet, pour n grand, $k_j \simeq \frac{t_j}{P_0} \simeq \frac{j\eta}{P_0}$, étant donné que l'espérance de deux temps consécutifs η existe. Ainsi $\sum_{j=1}^n (\frac{k_j}{n} - \frac{1}{2})^2 g'(\tilde{t}_j)^2$ devient $\sum_{j=1}^n (\frac{j\eta}{nP_0} - \frac{1}{2P_0})^2 g'(\tilde{t}_j)^2 = n \sum_{j=1}^n (\frac{j\eta}{nP_0} - \frac{1}{2P_0})^2 g'(\tilde{t}_j)^2 \frac{1}{n}$. Par la périodicité de g' , et que que $\frac{j}{n} \rightarrow u$, avec $0 < u < 1$, cette dernière somme converge vers

$$n \frac{\eta^2}{P_0^2} \int_0^1 \left(u - \frac{1}{2}\right)^2 du \int_0^1 g'(uP_0)^2 du + O(n^{-1}) = n \frac{\eta^2}{P_0^2} \frac{1}{12} \int_0^{P_0} g'(v)^2 dv \frac{1}{P_0} + O(n^{-1}), \quad (4.122)$$

avec $\int_0^1 (u - \frac{1}{2})^2 du = \frac{1}{12}$ et le changement de variable $v = uP_0$.

A présent calculons la variance de (4.120). Avec les mêmes arguments que pour le dénominateur de (4.120), la variance $\text{Var} [\nabla\text{SC}(P, g)]$ vaut

$$\text{Var} [n(P - P_0)] = \frac{12^2 P_0^6}{n^2 \eta^4 \left(\int_0^{P_0} g'(v)^2 dv \right)^2} \text{Cov} \left(\frac{1}{P_0} \sum_{j=1}^n \left(\frac{t_j}{n} - \frac{1}{2} \eta \right) g'(\tilde{t}_j) \varepsilon_j, \frac{1}{P_0} \sum_{k=1}^n \left(\frac{t_k}{n} - \frac{1}{2} \eta \right) g'(\tilde{t}_k) \varepsilon_k \right) + o(n^{-1}) \quad (4.123)$$

Calculons la covariance apparaissant dans (4.123), en utilisant que $g'(\tilde{t}_j) = g'(t_j)$.

$$\begin{aligned} \text{Cov} \left(\frac{1}{P_0} \sum_{j=1}^n \left(\frac{t_j}{n} - \frac{1}{2} \eta \right) g'(t_j) \varepsilon_j, \frac{1}{P_0} \sum_{k=1}^n \left(\frac{t_k}{n} - \frac{1}{2} \eta \right) g'(t_k) \varepsilon_k \right) \\ = \left(\frac{1}{P_0} \right)^2 \sum_{j=1}^n \sum_{k=1}^n \left(\frac{t_j}{n} - \frac{1}{2} \eta \right) g'(t_j) \left(\frac{t_k}{n} - \frac{1}{2} \eta \right) g'(t_k) \text{Cov}(\varepsilon_j, \varepsilon_k) \\ = \left(\frac{1}{P_0} \right)^2 \sum_{j=1}^n \sum_{k=1}^n \left(\frac{t_j}{n} - \frac{1}{2} \eta \right) g'(t_j) \left(\frac{t_k}{n} - \frac{1}{2} \eta \right) g'(t_k) \gamma(t_j - t_k), \end{aligned}$$

où $\gamma(t_j - t_k)$ est l'autocovariance du processus $\varepsilon(t)$ observé aux temps t_j et t_k . Par la relation entre la covariance $\gamma(s)$, $s \in \mathbb{R}$, et la densité spectrale $f_{sp}(\omega)$, $\omega \in \mathbb{R}$, du processus $\varepsilon(t)$, on remplace $\gamma(t_j - t_k)$ par $\frac{1}{2\pi} \int_{-\infty}^{+\infty} f_{sp}(\omega) \exp(-i\omega(t_j - t_k)) d\omega$. L'équation (4.124) devient

$$\frac{1}{P_0^2} \frac{1}{2\pi} \int_{-\infty}^{+\infty} \sum_{j=1}^n \sum_{k=1}^n \left(\frac{t_j}{n} - \frac{1}{2} \eta \right) g'(t_j) \left(\frac{t_k}{n} - \frac{1}{2} \eta \right) g'(t_k) f_{sp}(\omega) \exp(-i\omega(t_j - t_k)) d\omega,$$

qui peut se réécrire comme

$$\begin{aligned} \frac{1}{P_0^2} \frac{1}{2\pi} \int_{-\infty}^{+\infty} \sum_{j=1}^n \sum_{k=1}^n \left(\frac{t_j}{n} - \frac{1}{2} \eta \right) g'(t_j) \exp(-i\omega t_j) \left(\frac{t_k}{n} - \frac{1}{2} \eta \right) g'(t_k) \exp(i\omega t_k) f_{sp}(\omega) d\omega = \\ n \frac{1}{P_0^2} \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left| \sum_{j=1}^n \left(\frac{t_j}{n} - \frac{1}{2} \eta \right) g'(t_j) \exp(-i\omega t_j) \right|^2 \frac{1}{n} f_{sp}(\omega) d\omega. \quad (4.124) \end{aligned}$$

Dû au fait que, lorsque $n \rightarrow \infty$, $t_j \simeq j\eta$, $t_k \simeq k\eta$, et que $0 < \frac{j}{n}, \frac{k}{n} < 1$, par la périodicité de g' et des fonctions trigonométriques $\exp(-i\omega t_j)$, $\exp(i\omega t_k)$, ainsi que les temps distribués sur $[0, P_0]$ après repliement en phase sont uniformément distribués sur cette intervalle avec fonction de densité $\frac{1}{P_0}$, (4.126) converge vers,

$$\frac{1}{P_0^3} \frac{1}{2\pi} \eta^2 \int_{-\infty}^{+\infty} \int_0^1 \int_0^1 \left(u - \frac{1}{2} \right) \left(v - \frac{1}{2} \right) g'(v) \exp(i\omega v) dudv |\mathcal{TF}\{g'\}(\omega)|^2 f_{sp}(\omega) d\omega, \quad (4.125)$$

où $\mathcal{TF}\{g'\}(\omega) = \int_0^{P_0} g'(u) \exp(-i\omega u) du$ est la transformée de Fourier de g' . Remarquons que

$$\begin{aligned} \int_0^1 \int_0^1 \left(u - \frac{1}{2} \right) \left(v - \frac{1}{2} \right) dudv &\simeq \int_0^1 \left(u - \frac{1}{2} \right)^2 du + \int_0^1 \left(u - \frac{1}{2} \right) du \int_0^1 \left(v - \frac{1}{2} \right) dv \\ &\simeq \int_0^1 \left(u - \frac{1}{2} \right)^2 du = \frac{1}{12}, \end{aligned}$$

car $\int_0^1 (u - \frac{1}{2}) du = [(1 - \frac{1}{2})^2 - (\frac{1}{2})^2] = 0$. Ainsi (4.125) vaut

$$n \frac{1}{12} \eta^2 \frac{1}{P_0^3} \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\mathcal{TF}\{g'\}(\omega)|^2 f_{sp}(\omega) d\omega = n \frac{1}{12} \eta^2 \frac{1}{P_0^3} \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\mathcal{TF}\{g'\}(\omega)|^2 f_{sp}(\omega) d\omega. \quad (4.126)$$

En remplaçant la covariance dans (4.123) par (4.126), on obtient

$$\text{Var}[n(P - P_0)] = 12 \frac{1}{n\eta^2} P_0^3 \frac{1}{2\pi} \frac{\int_{-\infty}^{+\infty} |\mathcal{TF}\{g'\}(\omega)|^2 f_{sp}(\omega) d\omega}{\left(\int_0^{P_0} g'(u)^2 du\right)^2}. \quad (4.127)$$

En considérant le développement limité d'ordre 1 de la fonction $h(P) = \frac{2\pi}{P}$, la variance de l'estimateur de la fréquence vaut

$$\text{Var}[n(P - P_0)] = 12 \frac{1}{n\eta^2} \omega_0 \frac{\int_{-\infty}^{+\infty} |\mathcal{TF}\{g'\}(\omega)|^2 f_{sp}(\omega) d\omega}{\left(\int_0^{P_0} g'(u)^2 du\right)^2}. \quad (4.128)$$

■

On peut remarquer que pour le cas du signal sinusoidal, il n'est pas possible d'appliquer cette formule directement. En effet, la transformée de Fourier $\mathcal{TF}\{g'\}(\omega)$ d'un signal sinusoidal $g(t) = A \cos(\omega_0 t) + B \sin(\omega_0 t) + \mu$ est composée de fonctions de Dirac $\delta(\omega - \omega_0)$, $\delta(\omega + \omega_0)$. La dérivée de ces fonctions est simplement $\delta'(\omega - \omega_0)$. Le module au carré de $\mathcal{TF}\{g'\}(\omega)$ donne des fonction de Dirac de degré deux. L'intégrale $\int_{-\infty}^{+\infty} |\mathcal{TF}\{g'\}(\omega)|^2 f_{sp}(\omega) d\omega$ n'est pas possible à calculer. Pour le calcul de la variance asymptotique de $\hat{\omega}_n^{\text{NP}}$ pour un signal sinusoidal, on peut se référer à la preuve du théorème 9 sur la variance asymptotique de $\hat{\omega}_n^{\text{AH}}$, avec un signal sinusoidal.

4.3 Echantillonnage irrégulier ne vérifiant pas la condition de couverture de la période

Dans cette section, l'échantillonnage est irrégulier et ne vérifiant pas la condition de la couverture de la période (voir chapitre 1.10 sur l'échantillonnage). Nous considérons une régression non paramétrique sur des observations avec un signal périodique $g \in L^2[0, P_0)$ de période P_0 comme signal sous-jacent. Considérons le modèle (2.1), défini au chapitre 2.

$$Y_i = g(t_i) + \varepsilon(t_i), \quad (4.129)$$

où $\varepsilon(t)$ est un processus stochastique en temps continu. Les estimateurs $\hat{\omega}_n^{\text{NP}}$ et \hat{g}_n minimise la somme des carrés (4.130)

$$\text{SC}(\theta = [\omega, g]) = \sum_{i=1}^n (y_i - g(t_i))^2, \quad (4.130)$$

où l'estimateur du signal \hat{g}_n est la fonction de régression. L'estimateur \hat{g}_n est à estimer à partir des données en effectuant un repliement en phase des données pour chaque fréquence ω à considérer (voir paragraphe (2.4) sur le repliement en phase). Comme nous l'avons vu au paragraphe (2.4), La somme des carrés des résidus présente des minimums locaux aux fractions de la fréquence ω_0 . C'est pourquoi nous considérons que l'estimateur $\hat{\omega}_n^{\text{NP}}$ se trouve au voisinage de ω_0 .

4.3.1 Distribution asymptotique de l'estimateur $\widehat{\omega}_n^{\text{NP}}$

Considérons le modèle d'échantillonnage

$$T_n = n\delta + U, \quad (4.131)$$

avec $\delta > 0$, $U \sim \mathcal{U}(-\beta, \beta)$ avec $\beta > 0$. Les temps sont ordonnés, ce qui implique que $\beta < \frac{\delta}{2}$. Posons $L \in \mathbb{N}$ tel que $g(t + L\delta) \simeq g(t)$ pour tout $t \in [0, P_0)$. On a $g(t + L\delta) = g(t)$ si δ est une fraction de P_0 . En effet, dans ce cas, on peut toujours trouver un entier $K \in \mathbb{N}$, le nombre de période observée, tel que $L\delta = KP_0$. Autrement, nous avons seulement que $L\delta \simeq KP_0$. Tout comme pour le théorème 11, le résultat du théorème 12 est argumenté de manière informel, en considérant la procédure du calcul variationnel, comme utilisé dans la preuve du théorème 10.

Théorème 12 *Soit le modèle (4.1), avec erreur $\varepsilon(t)$, un processus stochastique Gaussien stationnaire et ergodique iid. La fonction $g(t)$ est périodique de période P_0 vérifiant les conditions (4.2) et (4.3). Soit $\widehat{g}(t)$ l'estimateur de la fonction g , et $\widehat{\omega}_n^{\text{NP}}$, l'estimateur de la fréquence ω_0 , définis à la section 2.4. Le modèle d'échantillonnage (4.131) est irrégulier et ne vérifie pas la condition de couverture de la période, définition 2, c'est-à-dire $\frac{P_0}{2L} > \beta > 0$. De plus, $\eta = \lim_{i \rightarrow \infty} \mathbb{E}[T_i - T_{i-1}] = \delta$. Soit $\widehat{\omega}_n^{\text{NP}}$ défini au point (2.21). Alors, lorsque $n \rightarrow \infty$, $n^{\frac{3}{2}}(\widehat{\omega}_n^{\text{NP}} - \omega_0)$ converge en distribution vers une loi Normale $\mathcal{N}(0, \tau^2)$, où*

$$\tau^2 = 2\pi \frac{L}{n^3} \omega_0 \frac{1}{\delta^2} 2\beta \frac{1}{\text{Int}_g} 12\sigma^2, \quad (4.132)$$

et

$$\text{Int}_g = \int_0^{\beta P_0} g'(u)^2 du + \sum_{l=1}^{L-1} \int_{l\delta_{P_0} P_0 - \beta P_0}^{l\delta_{P_0} + \beta P_0} g'(u)^2 du + \int_{P_0 - \beta P_0}^{P_0} g'(u)^2 du,$$

avec $\delta_{P_0} = \frac{P_0}{L}$.

Preuve. La preuve de ce théorème est presque identique au début de la preuve du théorème 11. La différence apparaît à l'équation (4.115), où l'intégrale $\int_0^{P_0} g(u)K'(\frac{t-u}{h})du$ n'est plus possible dans notre contexte de temps d'observation. Considérons de plus près cet échantillonnage. En moyenne, à la l -ème observation, on est au temps $l\delta$. C'est comme si on considérait des temps réguliers avec espacement δ . Alors, après L observations (ou K cycles du signal), on a une répétition des observations. En effet, les $L + 1$ à $2L$ observations suivantes sont identiques respectivement aux L premières. De plus la distance, en moyenne, entre chaque observation dans l'intervalle où se trouvent tous les autres intervalles superposés vaut $\delta_{P_0} = \frac{P_0}{L}$. Lorsque on effectue le repliement en phase, on obtient un ensemble de valeurs couvrant les intervalles $[l\delta_{P_0} - \beta P_0, l\delta_{P_0} + \beta P_0]$ avec $l = 1, \dots, L - 1$, $[0, \beta P_0]$ et $[P_0 - \beta P_0, P_0]$. Une des hypothèses du théorème est que $\beta < \frac{1}{2}\delta_{P_0}$. En effet, si $\beta \geq \frac{1}{2}\delta_{P_0}$, alors on retrouve un cas qui vérifie la condition de couverture de la période. Alors l'intégrale s'écrit comme

$$\begin{aligned} \int_0^{\beta P_0} g(u)K'(\frac{t-u}{h})du + \sum_{l=1}^{L-1} \int_{l\delta_{P_0} P_0 - \beta P_0}^{l\delta_{P_0} + \beta P_0} g(u)K'(\frac{t-u}{h})du + \int_{P_0 - \beta P_0}^{P_0} g(u)K'(\frac{t-u}{h})du \\ = \frac{1}{2} \frac{1}{P_0} g'(t) + O(h^2 + (nh)^{-\frac{1}{2}} n^{\xi_1}), \end{aligned} \quad (4.133)$$

pour tout $t \in [2h, \beta P_0]$, ou $t \in [l\delta_{P_0} P_0 - \beta P_0, l\delta_{P_0} + \beta P_0]$, $l = 1, \dots, L - 1$ ou encore $t \in [P_0 - \beta P_0, P_0 - 2h]$. A présent considérons la variance du résultat (4.120) apparaissant dans la preuve du théorème 11 .

$$\begin{aligned}
\text{Var}[n(P - P_0)] &= \text{Var} \left[\frac{\sum_{j=1}^n (\frac{k_j}{n} - \frac{1}{2}) g'(\tilde{t}_j) \varepsilon_j}{\sum_{j=1}^n (\frac{k_j}{n} - \frac{1}{2})^2 g'(\tilde{t}_j)^2} \right] + o(n^{-1}) \\
&= \left(\sum_{j=1}^n (\frac{k_j}{n} - \frac{1}{2})^2 g'(\tilde{t}_j)^2 \right)^{-1} \sigma^2 + o(n^{-1}). \tag{4.134}
\end{aligned}$$

Appliquons le modèle de temps (4.131) au temps d'observations. On a que $t_j = j \frac{P_0}{2L} + u_j$, avec u_j une réalisation de la variable aléatoire $U_j \sim \mathcal{U}(-\beta, \beta)$ et $0 < \beta < \frac{P_0}{L}$. Ainsi, pour n assez grand, $k_j \simeq \frac{t_j}{P_0} \simeq \frac{j\delta}{P_0} = j \frac{\delta}{P_0}$. Tout comme dans la preuve du théorème 11, dû au fait que, lorsque $n \rightarrow \infty$, $t_j \simeq j\eta$, $t_k \simeq k\eta$, et que $0 < \frac{j}{n}, \frac{k}{n} < 1$, par la périodicité de g' , ainsi que les temps distribués sur les intervalles $[0, \beta P_0]$, $[l\delta_{P_0} P_0 - \beta P_0, l\delta_{P_0} P_0 + \beta P_0]$, $l = 1, \dots, L-1$, et $[P_0 - \beta P_0, P_0]$, après repliement en phase sont uniformément distribués sur ces intervalles avec fonctions de densité $\frac{1}{2\beta P_0}$. En effet, on peut considérer les intervalles $[0, \beta P_0]$ et $[P_0 - \beta P_0, P_0]$ comme un intervalle $[-\beta P_0, \beta P_0]$ de longueur $2\beta P_0$. Nous obtenons que (4.134) converge vers

$$\frac{L P_0^2}{n \delta^2} \left(\int_0^1 \left(u - \frac{1}{2} \right)^2 du \frac{1}{2\beta P_0} \text{Int}_g \right)^{-1} \sigma^2 = \frac{L 24\beta P_0^3}{n \delta^2} \frac{1}{\text{Int}_g} \sigma^2,$$

où

$$\text{Int}_g = \sum_{l=1}^{L-1} \int_{l\delta_{P_0} P_0 - \beta P_0}^{l\delta_{P_0} P_0 + \beta P_0} g'(u)^2 du + \int_0^{\beta P_0} g'(u)^2 du + \int_{P_0 - \beta P_0}^{P_0} g'(u)^2 du.$$

En considérons le développement limité d'ordre 1 de la fonction $h(P) = \frac{2\pi}{P}$, la variance de l'estimateur de la fréquence vaut

$$2\pi \frac{L}{n^3} \omega_0 \frac{1}{\delta^2} 24\beta \frac{1}{\text{Int}_g} \sigma^2.$$

■

4.3.2 Cas particulier de ce modèle d'échantillonnage

Si $\beta = \frac{1}{2L}$, alors ce modèle d'échantillonnage donne la même variance asymptotique que celui obtenu par l'estimateur $\hat{\omega}_n^{\text{NP}}$ dans le cas d'un modèle d'échantillonnage vérifiant la condition de couverture de la période avec $\eta = \delta$. En effet, il suffit de remplacer $\beta = \frac{1}{2L}$ dans la variance asymptotique 4.132 et remarquer que $\text{Int}_g = \int_0^{P_0} g'(u)^2 du$. Si $\beta = 0$, alors l'échantillonnage est régulier d'un pas δ , dans ce cas la variance asymptotique 4.132 devient

$$\tau_{\text{NP}}^2 = L\omega_0^2 \frac{1}{\delta^2} \frac{1}{\sum_{l=0}^{L-1} g'(l\Delta_{P_0})^2} 12\sigma^2.$$

En effet, $\text{Int}_g = \text{Int}_g(\beta)$ et donc

$$\lim_{\beta \rightarrow 0} 2\pi \frac{L}{n^3} \omega_0 \frac{1}{\delta^2} 2\beta \frac{1}{\text{Int}_g} 3\sigma^2.$$

On voit que $\lim_{\beta \rightarrow 0} \tau_{\text{NP}}^2(\beta)$ est indéfini, car β et $\text{Int}_g(\beta)$ tendent les deux vers 0. On peut utiliser la règle de l'Hopital. Considérons seulement le rapport $\frac{2\beta}{\text{Int}_g}$. On obtient

$$\begin{aligned} \lim_{\beta \rightarrow 0} \frac{2\beta}{\text{Int}_g(\beta)} &= \lim_{\beta \rightarrow 0} \frac{2}{\frac{\partial}{\partial \beta} \text{Int}_g(\beta)} \\ &= \frac{1}{\sum_{l=0}^{L-1} g'(l\delta_{P_0})^2 P_0}, \end{aligned}$$

car

$$\begin{aligned} \lim_{\beta \rightarrow 0} \frac{\partial}{\partial \beta} \text{Int}_g(\beta) &= \lim_{\beta \rightarrow 0} \sum_{l=0}^{L-1} (g'(l\delta_{P_0} + \beta P_0)^2 + g'(l\delta_{P_0} P_0 - \beta P_0)^2) P_0 \\ &= 2 \sum_{l=0}^{L-1} g'(l\delta_{P_0})^2 P_0. \end{aligned}$$

On obtient ainsi le résultat. Si la fonction est sinuoidale, $g(t) = A \cos(\omega_0 t) + B \sin(\omega_0 t) + \mu$, alors $\sum_{l=0}^{L-1} g'(l\delta_{P_0})^2 = \frac{1}{2} L(A^2 + B^2) \omega_0^2$. Ainsi,

$$\tau_{\text{NP}}^2 = \frac{24\sigma^2}{\delta^2} \frac{1}{A^2 + B^2}.$$

Pour pouvoir considérer le cas $\beta = 0$, le pas δ doit vérifier la condition de Nyquist, $\delta \leq \frac{P_0}{2}$. Cette condition revient à dire que $L \geq 2$. En effet, la condition limite impose $L \frac{P_0}{2} = K P_0$, qui est équivalente à $L = 2K$, qui est en particulier vérifiée pour $K = 1$ et $L = 2$.

4.4 Exemples

Dans cette section, nous nous intéressons à la variance asymptotique de $\hat{\omega}^{\text{NP}}$, lorsque le signal est de type boîte. Nous donnons aussi quelques exemples de densités spectrales du processus Gaussien, pour illustrer l'influence de la corrélation des erreurs sur le comportement de la variance asymptotique.

4.4.1 Variance asymptotique de $\hat{\omega}_n^{\text{NP}}$ pour un signal boîte

Nous considérons le signal

$$g(t) = \begin{cases} 0 & 0 \leq t < t_0 - l \\ -R & t_0 - l \leq t < t_0 + l \\ 0 & t_0 + l \leq t < P_0 \end{cases}, \quad (4.135)$$

représenté à la figure 4.2. Nous considérons un échantillonnage de Poisson $\mathcal{P}(1)$. Etant donné que la dérivée d'une fonction constante par morceau donne des fonctions de Dirac, il n'est pas possible de calculer directement l'intégrale $\int_0^{P_0} g'(t)^2 dt$. De plus, pour pouvoir calculer la variance asymptotique, on tient compte de la méthode du repliement en phase des observations. De part le type d'échantillonnage considéré, l'intervalle de la période est couvert de façon uniforme. La moyenne des distances entre chaque observation successive diminue avec un ordre de $\frac{P_0}{n}$, lorsque $n \rightarrow \infty$. Ainsi, nous considérons la fonction \tilde{g}^n , avec n assez grand, approximant g et telle que $\tilde{g}^n \rightarrow g$ lorsque $n \rightarrow \infty$. Par n assez grand, on entend que $t_{11} - t_{12} = \frac{P_0}{n}$ en moyenne, avec $t_{11} < t_0 - l < t_{12}$ où t_{11} est le plus grand temps observé inférieur à $t_0 - l$ et t_{12} le plus petit temps observé supérieur à $t_0 - l$. On définit de la même façon t_{21} et t_{22} autour de $t_0 + l$,

$$\tilde{g}_n(t) = \begin{cases} 0 & 0 \leq t < t_{11} \\ R \frac{(t-t_{11})}{t_{12}-t_{11}} & t_{11} \leq t < t_{12} \\ -R & t_{12} \leq t < t_{21} \\ R \left(1 - \frac{(t-t_{21})}{t_{22}-t_{21}}\right) & t_{21} \leq t < t_{22} \\ 0 & t_{22} \leq t < P_0 \end{cases} .$$

La première dérivée de $\tilde{g}_n(t)$ vaut

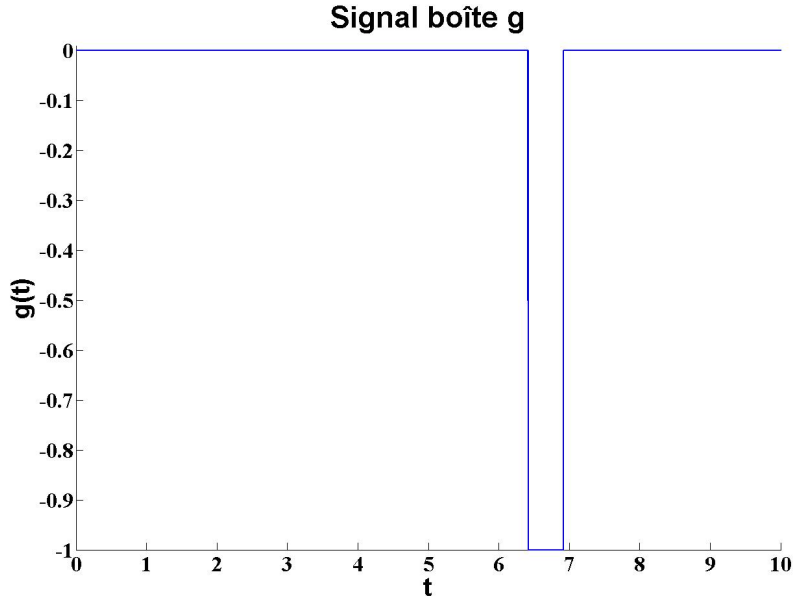


FIG. 4.2 – Signal boîte $g(t)$ contenant deux discontinuités

$$\tilde{g}_n(t) = \begin{cases} 0 & 0 \leq t < t_{11} \\ R \frac{n}{P_0} & t_{11} \leq t < t_{12} \\ 0 & t_{12} \leq t < t_{21} \\ -R \frac{n}{P_0} & t_{21} \leq t < t_{22} \\ 0 & t_{22} \leq t < P_0 \end{cases} .$$

La formule de la variance devient donc

$$\frac{1}{n^4} \tau^2 \simeq (2\pi)^2 \frac{3}{2} \frac{\sigma^2}{R^2} \frac{1}{n^4}, \quad (4.136)$$

car $\int_0^{P_0} \tilde{g}_n(t)^2 dt \simeq 2R^2 \frac{n^2}{P_0^2} \frac{P_0}{n} = 2R^2 \frac{n}{P_0}$. Cette formule donne une variance d'ordre n^{-4} .

Le tableau 4.1 donne les divers cas de figures où la distribution asymptotique de l'estimateur de la fréquence a été dérivé, ainsi que leurs auteurs.

4.4.2 Varier la densité spectrale

Soit la densité spectrale

Année	Auteurs	Méthode	Signal	Echantillonnage	Erreurs	Variance asymptotique
1971	Walker	AH	$A \cos(\omega_0 t) + B \sin(\omega_0 t) + \mu$	régulier	iid	$\frac{1}{n^3} \frac{24}{A^2+B^2} \sigma^2$
1973	Walker	AH	$A \cos(\omega_0 t) + B \sin(\omega_0 t) + \mu$	régulier	corrélées	$\frac{1}{n^3} \frac{24}{A^2+B^2} f_{sp}(\omega_0)$
2000	Hall, Reimann, Rice	NP	$g(t)$ vérifiant (4.2), (4.3)	irrégulier vérifiant déf. 2	iid	$\frac{1}{n^3} 24\pi\sigma^2\omega_0\eta^{-2} \left\{ \int_0^{P_0} g'(u)^2 du \right\}^{-1}$
2012	Nicoletti	AH	$g(t)$ vérifiant (4.2), (4.3), (4.4)	irrégulier vérifiant déf. 2	corrélées	$\frac{1}{n^3} \frac{24}{A_g^2+B_g^2} f_{sp}(\omega_0)$
2012	Nicoletti	NP	$g(t)$ vérifiant (4.2), (4.3)	irrégulier vérifiant la définition 2	corrélées	$\frac{12}{n^3} \omega_0 \frac{1}{\eta^2} \frac{\int_{-\infty}^{\infty} \mathcal{TF}\{g'\}(\omega) ^2 f_{sp}(\omega) d\omega}{\left(\int_0^{P_0} g'(u)^2 du\right)^2}$
2012	Nicoletti	régression avec g connu	$g(t)$ vérifiant (4.2), (4.3)	irrégulier vérifiant déf. 2	corrélées	$\frac{3}{n^3} \omega_0 \frac{1}{\eta^2} \frac{\int_{-\infty}^{\infty} \mathcal{TF}\{g'\}(\omega) ^2 f_{sp}(\omega) d\omega}{\left(\int_0^{P_0} g'(u)^2 du\right)^2}$
2012	Nicoletti	NP	$g(t)$ vérifiant (4.2), (4.3)	irrégulier ne vérifiant pas déf. 2, modèle (4.131)	iid	$2\pi \frac{L}{n^3} \omega_0 \frac{1}{\delta^2} 24\beta \frac{1}{Int_g} \sigma^2$
2012	Nicoletti	régression avec g connu	$g(t)$ signal boîte (4.135)	irrégulier vérifiant déf. 2	iid	$(2\pi)^2 \frac{3}{2} \frac{\sigma^2}{R^2} \frac{1}{n^4}$

TAB. 4.1 – Diverses configurations de la variance asymptotique de l'estimateur de la fréquence ainsi que leurs auteurs

$$f_{sp}(\omega) = \sigma^2 \frac{1}{\omega^2 + \phi^2}, \quad (4.137)$$

d'un processus CAR(1) de paramètre $\phi > 0$ (illustrations à la figure 4.3), défini par l'équation différentielle de premier ordre

$$\frac{d}{dt}\varepsilon(t) + \phi\varepsilon(t) = \frac{d}{dt}\varepsilon^*(t), \quad (4.138)$$

où $\varepsilon^*(t)$ est un processus stochastique continu Gaussien indépendant. La solution de (4.138) vaut (on peut voir par exemple [4], p. 298)

$$\varepsilon(t) = \int_0^{+\infty} \exp(-\phi v) d\varepsilon^*(t-v). \quad (4.139)$$

Le processus est stationnaire si $\phi > 0$.

Observons le comportement de $\int_{-\infty}^{+\infty} |\mathcal{TF}\{g'\}(\omega)|^2 f_{sp}(\omega) d\omega$, pour une fonction g donnée de fréquence ω_0 . Lorsque $0 < \phi < 1$, la densité spectrale (4.137) vérifie l'inégalité $f_{sp}(\omega) > \sigma^2$ pour les fréquences $|\omega| < \pm\sqrt{1-\phi^2}$. Si la transformée de Fourier $\mathcal{TF}\{g'\}(\omega)$ est à support compact inclus dans l'intervalle $[-\sqrt{1-\phi^2}, \sqrt{1-\phi^2}]$, alors la fréquence est moins bien estimée que dans le cas où les erreurs sont indépendantes. En effet,

$$\int_{-\infty}^{+\infty} |\mathcal{TF}\{g'\}(\omega)|^2 f_{sp}(\omega) d\omega > \int_{-\infty}^{+\infty} |\mathcal{TF}\{g'\}(\omega)|^2 d\omega = 2\pi \int_0^{P_0} g'(u)^2 du.$$

Ce qui a pour effet d'augmenter la variance asymptotique. Par contre, lorsque $\phi > 1$, toutes les fréquences $\omega > 0$ sont mieux estimées par rapport au cas où les erreurs sont iid. On remarque que pour deux fréquences $\omega_1, \omega_2 > 0$, telles que $\omega_1 < \omega_2$, la fréquence ω_2 aura une variance asymptotique inférieure à la fréquence ω_1 . Les simulations des effets de la densité spectrale sur la variance des estimateurs sont présentées à la section 5.3.

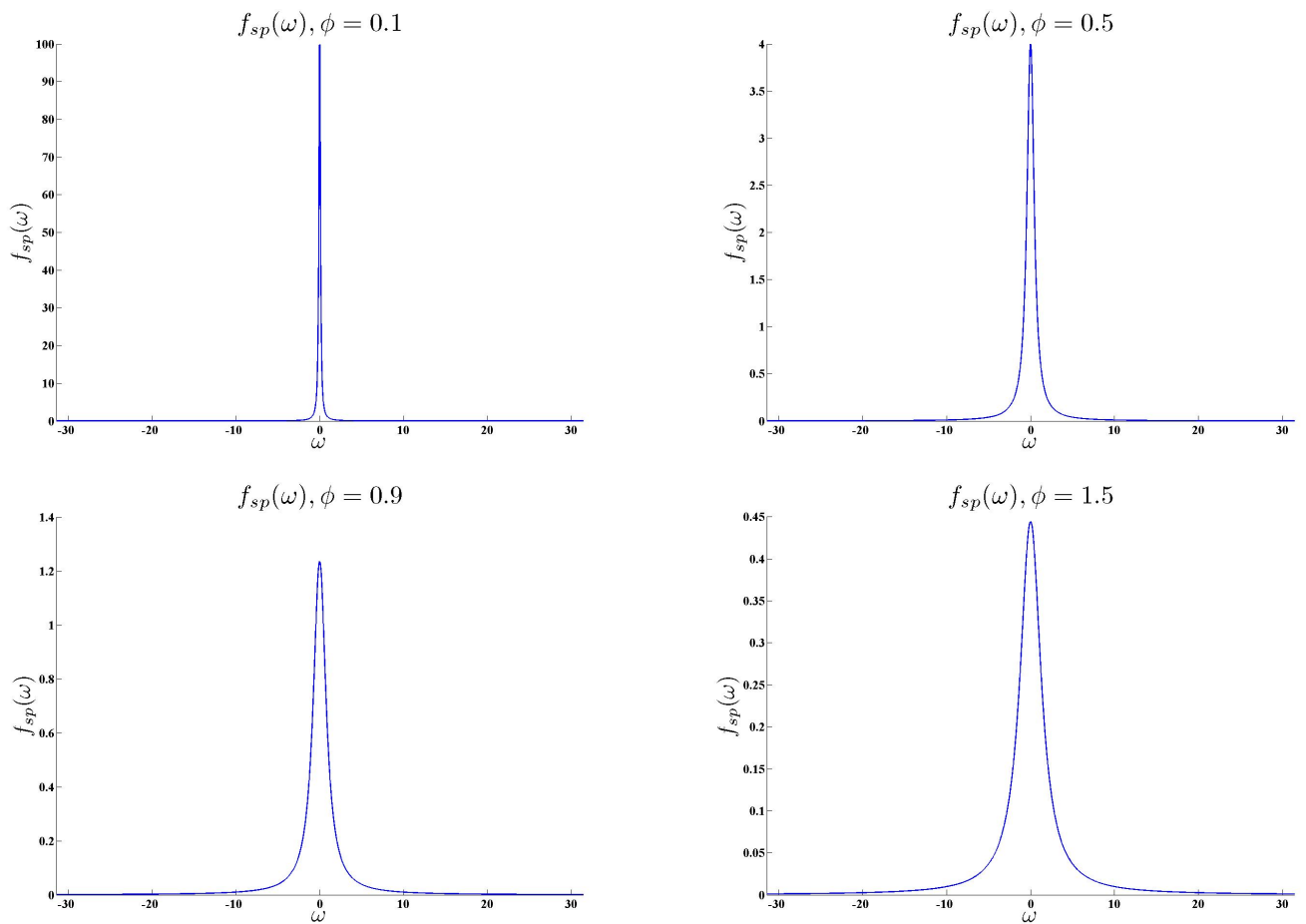


FIG. 4.3 – Densités spectrales $f_{sp}(\omega)$ d'un processus CAR(1) avec $\phi = 0.1, 0.5, 0.9, 1.5$, et $\sigma^2 = 1$.

Chapitre 5

Simulations

Dans ce chapitre, les méthodes présentées au chapitre 2 sont appliquées sur des données simulées. On considère plusieurs fonctions périodiques, avec différentes fréquences et amplitudes. Un processus Gaussien stationnaire indépendant représente les erreurs additives dans les données. Premièrement, nous considérons différents modèles d'échantillonnage vérifiant la condition de couverture de la période. Dans un deuxième temps, quelques modèles ne vérifiant pas la condition de couverture de la période seront considérés. Ensuite, une section est dédiée à la vérification du théorème 11, lorsque les erreurs suivent un processus Gaussien corrélé. L'avant-dernière section traite un signal spécifique comportant des discontinuités. La dernière section est consacrée à la comparaison de la variance asymptotique avec la variance estimée lorsque la fonction de régression est connue.

5.1 Modèles des données simulées

On considère le modèle des observations (2.1)

$$y_i = g_i + \varepsilon_i, i = 1, \dots, n,$$

où $g_i = g(t_i)$ est la fonction périodique g , de période P_0 , observée aux temps t_i . Le processus stochastique en temps discret ε_i est un processus Gaussien stationnaire et ergodique iid, un bruit blanc.

5.1.1 Signaux périodiques

Soit la fonction périodique de période P_0

$$g(t) = \sum_{k=0}^{\infty} g_i(t - kP_0)I_k(t). \quad (5.1)$$

où

$$I_k(t) = \begin{cases} 1 & \text{si } kP_0 \leq t < (k+1)P_0 \\ 0 & \text{sinon} \end{cases}. \quad (5.2)$$

Les fonctions $g_i(t)$, $i = 1, \dots, 6$, que nous considérons sont les suivantes

$$g_1(t) = A \cos(\omega_0 t) + B \sin(\omega_0 t) + \mu, \text{ avec } A, B, \mu \in \mathbb{R} \text{ et } R = \sqrt{A^2 + B^2} \text{ son amplitude.}$$

$$g_2(t) = \begin{cases} \frac{R}{C} (at^3 + bt^2 + ct + d) & 0 \leq t < \frac{P_0}{2} \\ R \left(1 - \frac{1}{C} (at^3 + bt^2 + ct + d)\right) & \frac{P_0}{2} \leq t < P_0, \end{cases},$$

où $C = a \left(\frac{P_0}{2}\right)^3 + b \left(\frac{P_0}{2}\right)^2 + c \left(\frac{P_0}{2}\right) + d$ et $R \in \mathbb{R}$. Ici nous avons choisi les valeurs $a = 10$, $b = 7$, $c = 3$ et $d = 0$.

$$g_3(t) = \begin{cases} \frac{R}{\exp\left(\frac{P_0}{4}\right)} \exp(t) & 0 \leq t < \frac{P_0}{4} \\ R \left(1 - \frac{4}{3P_0} \left(t - \frac{P_0}{4}\right)^2\right) & \frac{P_0}{4} \leq t < P_0 \end{cases},$$

où $R \in \mathbb{R}$.

$$g_4(t) = \begin{cases} Rc_1 & \text{si } 0 \leq t < \frac{P_0}{10} \\ Rc_2 & \text{si } \frac{P_0}{10} \leq t < \frac{P_0}{5} \\ Rc_3 & \text{si } \frac{P_0}{5} \leq t < \frac{2P_0}{5} \\ Rc_4 & \text{si } \frac{2P_0}{5} \leq t < \frac{3P_0}{5} \\ Rc_5 & \text{si } \frac{3P_0}{5} \leq t < \frac{9P_0}{10} \\ Rc_6 & \text{si } \frac{9P_0}{10} \leq t < P_0, \end{cases} \quad (5.3)$$

où, $c_i \in \mathbb{R}$ sont des constantes telles que le signal contient une périodicité interne au signal et $R > 0$. Ici nous avons choisi les valeurs $c_1 = 0.8$, $c_2 = -1$, $c_3 = -0.1$, $c_4 = 0.3$, $c_5 = -0.3$, $c_6 = 1$.

$$g_5(t) = \begin{cases} 0 & 0 \leq t < t_0 - l \\ -R & t_0 - l \leq t \leq t_0 + l \\ 0 & t_0 + l \leq t < P_0 \end{cases},$$

où t_0 est le centre de la partie non nulle du signal (ou boîte) et l est la demi-longueur de la partie non nulle du signal, $R > 0$. Ici nous avons choisi $t_0 = \frac{2P_0}{3}$ et $l = \frac{P_0}{40}$.

$$g_6(t) = R \left(\frac{2}{P_0} t + \frac{2}{\pi} \sin\left(2\pi \frac{t}{P_0}\right) - 1 \right), \quad (5.4)$$

où R est son amplitude.

L'amplitude des signaux g_1 et g_6 vaut R . Tandis que pour les signaux g_2 , g_3 , g_4 et g_5 , la notion d'amplitude est ambiguë. En effet, l'amplitude mesure la magnitude de la variation d'une fonction par rapport à sa valeur moyenne. Les signaux g_1 et g_6 ont une valeur moyenne égale à 0. Les signaux g_2 , g_3 , g_4 et g_5 , n'ont pas de symétrie autour de leur valeur moyenne. Pour les six fonctions, nous déterminerons l'amplitude de la fonction par R . On peut aussi parler d'amplitude pointe à pointe. Dans ce cas, les signaux g_1 et g_6 ont une amplitude $2R$, alors que les signaux g_2 , g_3 , g_4 et g_5 ont une amplitude R . Ici nous considérons la première définition. Les fonctions g_i , $i = 1, \dots, 6$ sont représentées dans la figure 5.1.

5.1.2 Modèles d'échantillonnage

Considérons des suites de temps t_1, \dots, t_n , avec des structures différentes, dont les modèles sont énumérés ci-dessous et avec $n = 100$ et 200 . Parmi les échantillonnages vérifiant la définition 2 de la couverture asymptotique de la période, nous considérons les modèles d'échantillonnage suivants :

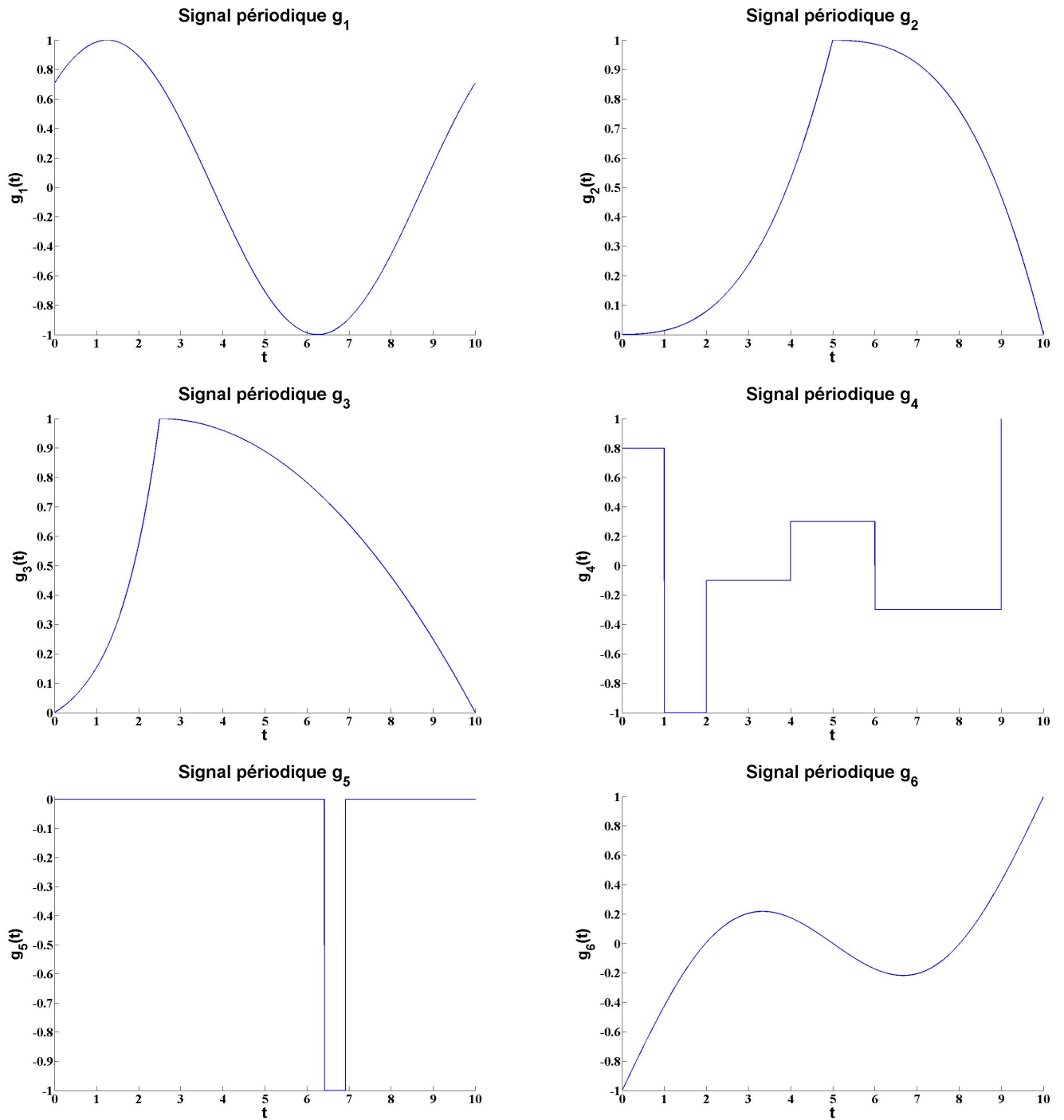


FIG. 5.1 – Signaux périodiques $g_1(t)$, $g_2(t)$, $g_3(t)$, $g_4(t)$, $g_5(t)$, $g_6(t)$ de gauche à droite, de haut en bas, respectivement avec $R = 1$.

- Ech₁** les temps uniformes ordonnés $t_i = nU_{(i)}$, où $U_{(i)}$ est la i -ème statistique d'ordre de l'ensemble de variables aléatoires $\{U_i\}_{i=1}^n$ avec $U_i \sim \mathcal{U}(0, 1)$;
- Ech₂** les temps correspondants aux temps des occurrences d'un processus de Poisson $\mathcal{P}(\lambda)$, $\lambda > 0$;
- Ech₃** les temps provenant d'un processus de renouvellement où les espacements entre chaque occurrence suit une loi de type mixture de deux lois exponentielles $p \exp(-\lambda_1) + (1-p) \exp(-\lambda_2)$, avec $\lambda_1, \lambda_2 > 0$ et $0 < p < 1$.
- Ech₄** les temps provenant d'un processus de type Neyman-Scott.

Mêmes si ces modèles d'échantillonnage ont la caractéristique que, asymptotiquement, les espacements entre deux temps successifs prennent toutes les valeurs possibles de l'axe réel, cela ne se vérifie plus pour un nombre fini d'observations. Ainsi, après un repliement en phase des données sur l'intervalle $[0, P_0)$, la faible densité des observations peut être assez faible, ce qui rend l'estimation de la forme de la fonction plus difficile. Le modèle de temps (5.5)

$$t_i = i\delta + U_i, \quad (5.5)$$

présenté dans la section 1.10 sur l'échantillonnage peut générer un processus de temps qui même asymptotiquement, ne peut pas toujours vérifier la condition de couverture de la période. Ainsi seulement une partie de la forme du signal peut être estimée.

5.2 Comparaison de la performances des méthodes : analyse harmonique, régression non paramétrique et périodogramme en amplitude (Deeming)

Dans cette section, nous procédons à une comparaison des performances des méthodes de l'analyse harmonique (AH), de la régression non paramétrique (NP) et du périodogramme en amplitude, appelée méthode de Deeming (D). Nous avons simulé $N = 1000$ fois des séries temporelles et effectué l'estimation. Les symboles $I_{n,i}^{\text{AH}}$, $I_{n,i}^{\text{NP}}$ et $I_{n,i}^{\text{D}}$, à coefficient 95%, pour ω_0 , définissent les intervalles de confiance basés sur la variance asymptotique τ_{AH} , ((4.76), théorème 9) et τ_{NP} ((4.110), théorème 11), dans le cas d'erreurs Gaussiennes iid. Les symboles $\hat{\omega}_{n,i}^{\text{AH}}$, $\hat{\omega}_{n,i}^{\text{NP}}$ et $\hat{\omega}_{n,i}^{\text{D}}$ décrivent les valeurs estimées trouvées lors de l' i -ème simulation. Nous utiliserons les symboles $\hat{\omega}_n^{\text{AH}}$, $\hat{\omega}_n^{\text{NP}}$ et $\hat{\omega}_n^{\text{D}}$, lorsque nous considérons l'estimateur et non une réalisation de celui-ci. Nous considérons la variance asymptotique τ_n^{AH} pour l'intervalle confiance I_n^{D} . Dans la suite, le symbole $\hat{\omega}_n$ représente aussi bien $\hat{\omega}_n^{\text{AH}}$, $\hat{\omega}_n^{\text{NP}}$ et $\hat{\omega}_n^{\text{D}}$. De même pour les intervalles de confiance estimés, où nous utiliserons le symbole $I_{n,i}$. Ainsi

$$I_{n,i} = [\hat{\omega}_{n,i} - 2\tau, \hat{\omega}_{n,i} + 2\tau],$$

où 2 est le quantile $\Phi^{-1}(0.95)$ de la loi normale $\mathcal{N}(0, 1)$, avec $\Phi(x)$ la fonction de distribution de la loi normale. Nous calculons aussi les estimateurs de la variance τ_n^{AH} et τ_n^{NP}

$$s_1^2 = \frac{1}{N_1 - 1} \sum_{i \in \mathcal{V}_1} (\hat{\omega}_{n,i} - \omega_0)^2,$$

où $\hat{\omega}_{n,i}$ est l'estimateur de la fréquence pour le i -ème échantillon d'observations et

$$\mathcal{V}_1 = \left\{ i \in \{1, \dots, n\} : \frac{3}{4}\omega_0 \leq \hat{\omega}_{n,i} \leq \frac{5}{4}\omega_0 \right\},$$

$N_1 = \text{card}(\mathcal{V}_1)$, la cardinalité de l'ensemble \mathcal{V}_1 . La restriction $i \in \mathcal{V}_1$ a été choisie pour ne pas considérer les mauvaises estimations, où les estimations d'un multiple de la période. Nous avons vu

dans le chapitre 3 sur les données Hipparcos, qu'un signal de type binaire à éclipse présente une période interne au signal qui induit la détection de la demi-période par les méthodes de Deeming et d'analyse harmonique. Nous considérons justement un signal de ce genre, le signal g_4 . Ainsi, pour le calcul de s^2 , il nous faut donc considérer la version

$$s_2^2 = \frac{1}{N_2 - 1} \sum_{i \in \mathcal{V}_2} (\hat{\omega}_{n,i} - 2\omega_0)^2,$$

pour les méthodes de Deeming et de l'analyse harmonique appliquées au signal g_4 .

$$\mathcal{V}_2 = \left\{ i \in \{1, \dots, n\} : \frac{3}{2}\omega_0 \leq \hat{\omega}_{n,i} \leq \frac{5}{2}\omega_0 \right\},$$

et $N_2 = \text{card}(\mathcal{V}_2)$. Nous utiliserons le symbole s pour caractériser l'estimateur de l'écart-type asymptotique, qui prendra les valeurs s_1 ou s_2 selon la méthode et le signal considérés, comme indiqué au tableau 5.1.

	g_1	g_2	g_3	g_4
D	s_1	s_1	s_1	s_2
AH	s_1	s_1	s_1	s_2
NP	s_1	s_1	s_1	s_1

TAB. 5.1 – Table des estimateurs de l'écart-type selon la méthode et le signal considéré.

Soit les ensembles

$$\mathcal{A}_{\text{AH}} = \{i \in \{1, \dots, N\} : \omega_0 \in I_{n,i}^{\text{AH}}\}, \mathcal{A}_{\text{NP}} = \{i \in \{1, \dots, N\} : \omega_0 \in I_{n,i}^{\text{NP}}\}, \mathcal{A}_{\text{D}} = \{i \in \{1, \dots, N\} : \omega_0 \in I_{n,i}^{\text{D}}\},$$

représentant les indices $i \in \{1, \dots, N\}$ tels que l'intervalle de confiance $\hat{I}_{n,i}$ couvre la vraie fréquence. Définissons $p_{\text{AH}} = \frac{\text{card}(\mathcal{A}_{\text{AH}})}{N}$, $p_{\text{NP}} = \frac{\text{card}(\mathcal{A}_{\text{NP}})}{N}$ et $p_{\text{D}} = \frac{\text{card}(\mathcal{A}_{\text{D}})}{N}$, les proportions des valeurs des estimateurs $\hat{\omega}_n^{\text{AH}}$, $\hat{\omega}_n^{\text{NP}}$ et $\hat{\omega}_n^{\text{D}}$, respectivement, considérées comme bien estimées selon le critère CC₂ (détection de la période ou de la demi-période, voir chapitre 3). Nous utilisons le symbole p , lorsqu'on ne se réfère pas à une des méthodes en particulier. Pour chaque modèle d'échantillonnage, nous effectuons les comparaisons sur les signaux g_1, \dots, g_4 , avec fréquences $\omega_0 \in \{0.1, 0.8, 1.5\}$ et un rapport signal bruit $\text{SNR} = \frac{R}{\sigma} \in \{1, 2, 3\}$.

Les résultats des comparaisons des performances des trois méthodes avec les modèles d'échantillonnages Ech₁, Ech₂, Ech₃ et Ech₄, vérifiant la condition de couverture de la période, définition 2, sont présentés dans les tableaux 5.3 à 5.6, où nous considérons uniquement le critère de comparaison CC₂. Les résultats des comparaisons des performances des trois méthodes avec les modèles d'échantillonnages EchNU₁ et EchNU₂, ne vérifiant pas la condition de couverture de la période, sont présentés dans les tableaux 5.3 à 5.6, où nous considérons uniquement le critère de comparaison CC₂.

La section 5.4 sera consacrée à l'étude de la fonction particulière $g_5(t)$. On s'intéresse à la variance de $\hat{\omega}_n^{\text{NP}}$. On verra que $\hat{\omega}_n^{\text{AH}}$ a une performance nettement inférieure. L'avant-dernière partie sera consacrée à la vérification du théorème 11 sur la variance asymptotique des estimateurs $\hat{\omega}_n^{\text{AH}}$ et $\hat{\omega}_n^{\text{NP}}$, dans le cas où les erreurs suivent un processus Gaussien stationnaire avec fonction d'autocovariance non nulle pour des décalages non nuls. La dernière section sera consacrée à l'étude de la variance asymptotique de l'estimateur de la fréquence $\hat{\omega}_n^{\text{NP}}$ pour les signaux g_1 à g_4 , avec un processus de temps vérifiant la condition de la couverture de la période 2. La fonction de régression sera considérée comme connue, ce qui permettra une bonne comparaison de l'écart-type s avec l'écart-type asymptotique. Nous effectuons aussi ces comparaisons avec deux modèles d'échantillonnages ne vérifiant pas la définition 2 de la condition de couverture de la période. Cette dernière partie était nécessaire pour pouvoir effectuer une bonne comparaison entre s et la variance théorique de l'estimateur $\hat{\omega}_n^{\text{NP}}$. En effet, nous verrons

que les estimations obtenues, lorsqu'on considère une fonction régression estimée avec $n \in \{100, 200\}$ données, ont souvent une grande variabilité, due au manque de précision de la forme de la fonction estimée.

5.2.1 Modèles d'échantillonnage avec la propriété de couverture de la période

Les comparaisons sont effectuées avec le critère CC₂. Une estimation de la fréquence $\hat{\omega}_{n,i}$ est considérée comme correcte si l'intervalle de confiance asymptotique $I_{n,i}$ contient la fréquence ω_0 , ou $2\omega_0$, dans le cas particulier du signal g_4 avec les méthodes D et AH. Elles sont représentées dans les tableaux 5.3 à 5.6, pour l'échantillonnage uniforme Ech₁, pour l'échantillonnage provenant d'un processus de Poisson Ech₂, pour l'échantillonnage provenant d'un processus de mixture d'exponentielles Ech₃ et pour l'échantillonnage provenant d'un processus de Neymann-Scott Ech₄. La comparaison est effectuée pour chaque signaux de fréquence $\omega_0 \in \{0.1, 0.8, 1.5\}$ et de rapport signal/bruit SNR $\in \{1, 2, 3\}$. Ces simulations sont chacune composée d'un nombre $n \in \{100, 200\}$ d'observations. Le tableau 5.2 contient les valeurs des écarts-types τ_{AH} et τ_{NP} . Pour le signal g_1 , l'intégrale de g'_1 vaut

$$\int_0^{P_0} (g'_1(t))^2 dt = (A^2 + B^2) \omega_0 \pi = R^2 \omega_0 \pi, \quad (5.6)$$

Pour SNR = 1 = $\frac{\sigma}{R}$, l'écart-type τ_{NP} vaut

$$\sqrt{2\pi 12 \frac{\omega_0}{\omega_0 \pi} \frac{1}{n^3}} = \sqrt{\frac{24}{n^3}} = \begin{cases} 0.0048 & \text{si } n = 100 \\ 0.0017 & \text{si } n = 200 \end{cases}.$$

Pour le signal g_2 , l'intégrale de g'_2 vaut

$$\begin{aligned} \int_0^{P_0} (g'_2(t))^2 dt &= 2 \frac{R^2}{C^2} \left(9a^2 \frac{1}{5} \left(\frac{P_0}{2} \right)^5 + 3ab \left(\frac{P_0}{2} \right)^4 + \left(2ac + \frac{4}{3}b^2 \right) \left(\frac{P_0}{2} \right)^3 \right. \\ &\quad \left. + 2bc \left(\frac{P_0}{2} \right)^2 + c^2 \left(\frac{P_0}{2} \right) \right) \\ &\simeq 0.113R^2, \end{aligned}$$

où $C = a \left(\frac{P_0}{2} \right)^3 + b \left(\frac{P_0}{2} \right)^2 + c \left(\frac{P_0}{2} \right) + d$. Pour SNR = 1 = $\frac{\sigma}{R}$, l'écart-type τ_{NP} vaut

$$\sqrt{2\pi 12 \frac{\omega_0}{0.113} \frac{1}{n^3}} = \begin{cases} 0.0081 & \text{si } n = 100 \\ 0.0029 & \text{si } n = 200 \end{cases},$$

Pour la fonction g_3 , on obtient

$$\begin{aligned} \int_0^{P_0} (g'_3(t))^2 dt &= R^2 \left(\frac{1}{2} \frac{\exp\left(\frac{P_0}{2}\right)}{\exp\left(\frac{P_0}{4}\right)} + P_0 \right) \\ &\simeq 0.528R^2. \end{aligned}$$

En considérant à nouveau $\frac{\sigma^2}{R^2} = 1$, on obtient l'écart-type τ_{NP} suivant :

$$\sqrt{2\pi 12 \frac{\omega_0}{0.528} \frac{1}{n^3}} = \begin{cases} 0.00377 & \text{si } n = 100 \\ 0.00133 & \text{si } n = 200 \end{cases}.$$

Pour la fonction g_4 , étant donné qu'il s'agit d'une fonction constante par morceau, qui contient des discontinuités, on se base sur les calculs effectués à la section 4.4.1 pour le signal g_5 . La fonction g_4 , définie au point (5.3), contient six discontinuités aux temps $t_1 = \frac{P_0}{10}$, $t_2 = \frac{P_0}{5}$, $t_3 = 2\frac{P_0}{5}$, $t_4 = 3\frac{P_0}{5}$, $t_5 = 9\frac{P_0}{10}$ et $t_6 = P_0$. Les sauts de la fonction aux temps $t_i, i = 1, \dots, 6$ ont des amplitudes de $\tilde{c}_1 = -1.8$, $\tilde{c}_2 = 0.9$, $\tilde{c}_3 = 0.4$, $\tilde{c}_4 = -0.6$, $\tilde{c}_5 = 1.3$ et $\tilde{c}_6 = -0.2$. Comme pour le signal g_5 , pour n assez grand, on a $t_{11} - t_{12} = \frac{P_0}{n}$ en moyenne, avec $t_{11} < t_1 < t_{12}$ où t_{11} est le plus grand temps observé inférieur à t_1 et t_{12} le plus petit temps observé supérieur à t_1 . On définit de la même façon t_{i1} et t_{i2} autour de $t_i, i = 2, \dots, 6$. Considérons la fonction

$$\tilde{g}_n(t) = \sum_{i=1}^6 \tilde{g}_{n,i}(t),$$

où $i = 1, \dots, 6$ et

$$\tilde{g}_{n,i}(t) = \begin{cases} 0 & 0 \leq t < t_{i1} \\ R \left(c_i + \tilde{c}_i \frac{(t-t_{i1})}{t_{i2}-t_{i1}} \right) & t_{i1} \leq t < t_{i2} \\ Rc_{i+1} & t_{i2} \leq t < t_{(i+1)1} \\ 0 & t_{(i+1)1} \leq t < P_0 \end{cases},$$

en posant $c_7 = c_1$, et R l'amplitude. La première dérivée de $\tilde{g}_n(t)$ donne

$$\frac{d}{dt} \tilde{g}_n(t) = \sum_{i=1}^6 \frac{d}{dt} \tilde{g}_{n,i}(t),$$

où

$$\frac{d}{dt} \tilde{g}_{n,i}(t) = \begin{cases} 0 & 0 \leq t < t_{i1} \\ R\tilde{c}_i \frac{n}{P_0} & t_{i1} \leq t < t_{i2} \\ 0 & t_{i2} \leq t < P_0 \end{cases}.$$

On peut donc approximer l'intégrale

$$\int_0^{P_0} (g'_4(t))^2 dt \simeq R^2 \sum_{i=1}^6 \tilde{c}_i^2 \frac{P_0}{n} = 6.3R^2 \frac{P_0}{n}.$$

Pour un SNR = 1, l'écart-type asymptotique τ_{NP} vaut

$$(2\pi)^2 \frac{12}{6.3} \frac{1}{n^4}.$$

Analyse des résultats pour l'échantillonnage Ech₁ (tableau 5.3)

La méthode AH est toujours supérieure aux autres méthodes pour les quatre signaux g_1, g_2, g_3 et g_4 . Ce pendant les performances sont relativement proches pour le signal g_4 lorsque SNR = 3. On peut comparer l'écart-type estimé $s_{AH,1}$, avec l'écart-type asymptotique τ_{AH} , tableau 5.2. La valeur de $s_{AH,1}$, avec $\omega_0 = 0.1, 0.8, 1.5, g_1$, SNR = 1, 2, 3 et $n = 100, 200$, est proche de la valeur τ_{AH} . On observe cela aussi pour g_4 avec $\omega_0 = 0.1, 0.8, 1.5$, SNR = 2, 3 et $n = 200$. Pour les signaux g_2 et g_3 , la convergence de $s_{AH,1}$ vers τ_{AH} semble plus lente, avec l'échantillonnage Ech₁. La valeur de s_{NP} est le plus souvent assez éloignée de la valeur de τ_{NP} , pour les quatre signaux, les trois fréquences ainsi que les trois valeurs du SNR considérés. Ici le choix de la taille du lisseur appliqué aux observations repliées en phase, pour estimer la fonction de régression, peut détériorer la précision de l'estimation.

Analyse des résultats pour l'échantillonnage Ech₂ (tableau 5.4)

Avec cet échantillonnage, les estimateurs de la fréquence $\hat{\omega}_n$ ont une variance asymptotique qui est égale à la variance asymptotique lorsque l'échantillonnage est Ech₁. En effet, dans ces deux modèles d'échantillonnage, vérifiant la condition 2, η , l'espérance de la différence entre deux temps aléatoires consécutifs, vaut 1. Pour $n = 100$, premier tableau de 5.4, les résultats sont en effet proches de ceux obtenus avec l'échantillonnage Ech₁. Pour $n = 200$, on observe que p_{NP} est supérieur à p_{AH} , pour la fonction g_2 , lorsque $n = 200$ et $\omega_0 = 1.5$, contrairement à l'échantillonnage Ech₁, où la différence entre les deux méthodes est minimale. On obtient les mêmes conclusions que celles de l'échantillonnage Ech₁, lors de la comparaison des écarts-type estimés avec les écarts-type asymptotiques.

Analyse des résultats pour l'échantillonnage Ech₃ (tableau 5.5)

Dans l'échantillonnage Ech₃, $\eta = p\frac{1}{\lambda_1} + (1-p)\frac{1}{\lambda_2} = 4.06$, avec $p = 0.6$, $\lambda_1 = 10$, $\lambda_2 = 0.1$. Si on compare les valeurs de $s_{\text{AH},j}$, $j = 1, 2$, $\omega_0 \in \{0.1, 0.8, 1.5\}$ et g_i , $i = 1, \dots, 4$, avec celles obtenues dans les tableaux 5.3 et 5.4, on remarque qu'il y a un facteur d'environ 4. Par exemple, au tableau 5.3, SNR = 1, signal g_1 , et $\omega_0 = 0.1$ on a que $s_{\text{AH},1} = 0.0055$. Dans le tableau (5.4) pour le même signal, SNR = 1 et $\omega_0 = 0.1$, on obtient $s_{\text{AH},1} = 0.0016$. Cet échantillonnage donne un net avantage à la méthode AH. Pour comparer les valeurs de s_{AH} et s_{NP} avec τ_{AH} et τ_{NP} , respectivement, il suffit de diviser par 4.06 les valeurs du tableau 5.2.

Analyse des résultats pour l'échantillonnage Ech₄ (tableau 5.6)

Cet échantillonnage présente la propriété de créer des groupements d'observations, dont ces groupes sont séparés par des distances suivant une loi exponentielle. Asymptotiquement, ce genre d'échantillonnage permet tout de même une couverture de la période. Cependant pour une taille finie d'échantillon, les intervalles sans observations peuvent poser des problèmes quant à l'estimation de la fréquence. Pour l'échantillonnage Ech₄, $\eta = p\frac{1}{2\lambda} + (1-p)\frac{1}{z}$, où $p = 0.1$, la proportion du nombre de distances inter-groupes et la valeur de $\frac{1}{2\lambda}$ est la distance moyenne inter-groupes. Le facteur $(1-p)$ est la proportion du nombre d'observations intra-groupes et z est la longueur moyenne de l'intervalle générant les temps aléatoires uniformes pour les temps intra-groupes. Nous avons $\lambda = 0.2$, $z = 2.5$. Ainsi $\eta = 0.610$. Pour comparer les valeurs de $s_{\text{AH},1}$, $s_{\text{AH},2}$ et s_{NP} avec τ_{AH} et τ_{NP} , respectivement, il suffit de diviser par 0.610 les valeurs du tableau 5.2. On remarque que la meilleure performance est atteinte par la méthode AH. Cela se vérifie surtout pour le signal g_1 . On observe que la performance est meilleure pour la fréquence $\omega_0 = 1.5$. Pour $n = 200$, $\omega_0 = 1.5$ et g_1 , SNR = 1, 2, 3, nous pouvons comparer les valeurs de $s_{\text{AH},1}$ avec $\frac{\tau_{\text{AH}}}{\eta}$.

		Ecart-type asymptotique, $n = 100$											
		SNR = 1				SNR = 2				SNR = 3			
		g_1	g_2	g_3	g_4	g_1	g_2	g_3	g_4	g_1	g_2	g_3	g_4
$\omega_0 = 0.1$	AH	48e-4	96e-4	96e-4	48e-4	24e-4	48e-4	48e-4	24e-4	16e-4	32e-4	32e-4	16e-4
	NP	48e-4	81e-4	37e-4	8e-4	24e-4	40e-4	18e-4	4e-4	16e-4	27e-4	12e-4	3e-4
$\omega_0 = 0.8$	AH	48e-4	96e-4	96e-4	48e-4	24e-4	48e-4	48e-4	24e-4	16e-4	32e-4	32e-4	16e-4
	NP	48e-4	81e-4	82e-4	8e-4	24e-4	40e-4	41e-4	4e-4	16e-4	27e-4	27e-4	3e-4
$\omega_0 = 1.5$	AH	48e-4	96e-4	96e-4	48e-4	24e-4	48e-4	48e-4	24e-4	16e-4	32e-4	32e-4	16e-4
	NP	48e-4	86e-4	87e-4	8e-4	24e-4	43e-4	43e-4	4e-4	16e-4	28e-4	29e-4	3e-4

		Ecart-type asymptotique, $n = 200$											
		SNR = 1				SNR = 2				SNR = 3			
		g_1	g_2	g_3	g_4	g_1	g_2	g_3	g_4	g_1	g_2	g_3	g_4
$\omega_0 = 0.1$	AH	17e-4	34e-4	34e-4	17e-4	9e-4	18e-4	18e-4	9e-4	6e-4	12e-4	12e-4	6e-4
	NP	17e-4	28e-4	13e-4	2e-4	9e-4	14e-4	6e-4	1e-4	6e-4	9e-4	4e-4	7e-05
$\omega_0 = 0.8$	AH	17e-4	34e-4	34e-4	17e-4	9e-4	18e-4	18e-4	9e-4	6e-4	12e-4	12e-4	6e-4
	NP	17e-4	29e-4	29e-4	2e-4	9e-4	15e-4	14e-4	1e-4	6e-4	9e-4	9e-4	7e-05
$\omega_0 = 1.5$	AH	17e-4	34e-4	34e-4	17e-4	9e-4	18e-4	18e-4	9e-4	6e-4	12e-4	12e-4	6e-4
	NP	17e-4	30e-4	31e-4	2e-4	9e-4	15e-4	15e-4	1e-4	6e-4	10e-4	10e-4	7e-05

TAB. 5.2 – Table des valeurs des écarts-type asymptotiques, τ pour les méthodes AH et NP et pour différentes fréquences ω_0 , signaux g et SNR, pour l'échantillonnage vérifiant la définition 2, et $\eta = 1$. L'échantillon est de taille $n = 100$, pour le premier tableau, et $n = 200$ pour le second.

		Ech₁, n = 100												
		SNR = 1				SNR = 2				SNR = 3				
		<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	
$\omega_0 = 0.1$	<i>p</i>	D	0.90	0.66	0.64	0.21	0.68	0.93	0.86	0.48	0.58	0.86	0.72	0.55
		AH	0.91	0.64	0.59	0.23	0.91	0.93	0.89	0.54	0.95	0.94	0.85	0.64
		NP	0.84	0.59	0.40	0.22	0.70	0.80	0.43	0.59	0.74	0.70	0.37	0.71
	<i>s</i>	D	56e-4	86e-4	82e-4	12e-3	42e-4	51e-4	61e-4	8e-4	39e-4	43e-4	56e-4	7e-3
		AH	55e-4	93e-4	94e-4	12e-3	27e-4	5e-3	58e-4	75e-4	18e-4	34e-4	45e-4	63e-4
		NP	72e-4	92e-4	85e-4	84e-4	42e-4	63e-4	58e-4	9e-3	33e-4	5e-3	48e-4	9e-3
$\omega_0 = 0.8$	<i>p</i>	D	0.91	0.63	0.41	0.28	0.81	0.92	0.83	0.55	0.76	0.90	0.81	0.63
		AH	0.91	0.613	0.40	0.30	0.91	0.93	0.87	0.61	0.95	0.94	0.87	0.71
		NP	0.89	0.54	0.40	0.25	0.77	0.82	0.79	0.56	0.79	0.72	0.71	0.67
	<i>s</i>	D	54e-4	27e-3	41e-3	7e-2	34e-4	53e-4	68e-4	11e-3	3e-3	40e-4	50e-4	62e-4
		AH	53e-4	27e-3	41e-3	71e-3	27e-4	50e-4	62e-4	10e-3	18e-4	34e-4	42e-4	55e-4
		NP	61e-4	14e-3	17e-3	16e-3	36e-4	65e-4	70e-4	80e-4	28e-4	48e-4	51e-4	81e-4
$\omega_0 = 1.5$	<i>p</i>	D	0.92	0.58	0.42	0.9	0.81	0.91	0.82	0.56	0.77	0.88	0.0	0.62
		AH	0.94	0.57	0.43	0.29	0.93	0.94	0.87	0.60	0.97	0.95	0.88	0.71
		NP	0.90	0.51	0.39	0.25	0.75	0.84	0.78	0.54	0.76	0.72	0.67	0.67
	<i>s</i>	D	53e-4	70e-3	94e-3	12e-2	34e-4	94e-4	19e-3	71e-4	29e-4	42e-4	50e-4	63e-4
		AH	50e-4	73e-3	96e-3	12e-2	25e-4	49e-4	11e-3	24e-3	17e-4	33e-4	42e-4	55e-4
		NP	61e-4	37e-3	46e-3	37e-3	38e-4	59e-4	77e-4	21e-3	29e-4	48e-4	54e-4	86e-4

		Ech₁, n = 200												
		SNR = 1				SNR = 2				SNR = 3				
		<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	
$\omega_0 = 0.1$	<i>p</i>	D	0.95	0.93	0.86	0.80	0.82	0.96	0.91	0.96	0.88	0.95	0.87	0.98
		AH	0.91	0.88	0.85	0.83	0.88	0.93	0.90	0.99	0.97	0.95	0.89	0.99
		NP	0.78	0.72	0.43	0.74	0.66	0.61	0.48	0.97	0.81	0.47	0.21	0.99
	<i>s</i>	D	18e-4	30e-4	36e-4	37e-4	12e-4	16e-4	21e-4	19e-4	10e-4	13e-4	17e-4	13e-4
		AH	20e-4	38e-4	39e-4	33e-4	10e-4	19e-4	21e-4	12e-4	7e-4	13e-4	16e-4	4e-4
		NP	29e-4	46e-4	42e-4	46e-4	16e-4	28e-4	23e-4	44e-4	12e-4	22e-4	18e-4	49e-4
$\omega_0 = 0.8$	<i>p</i>	D	0.93	0.90	0.73	0.79	0.78	0.94	0.87	0.96	0.86	0.92	0.87	0.98
		AH	0.95	0.89	0.73	0.82	0.92	0.96	0.91	0.98	0.99	0.98	0.92	0.99
		NP	0.85	0.72	0.59	0.70	0.70	0.68	0.64	0.94	0.84	0.58	0.56	0.97
	<i>s</i>	D	20e-4	95e-4	19e-3	23e-3	13e-4	18e-4	23e-4	20e-4	11e-4	14e-4	17e-4	13e-4
		AH	18e-4	74e-4	17e-3	22e-3	9e-4	17e-4	21e-4	14e-4	6e-4	11e-4	15e-4	7e-4
		NP	25e-4	62e-4	78e-4	86e-4	14e-4	26e-4	29e-4	42e-4	11e-4	19e-4	19e-4	42e-4
$\omega_0 = 1.5$	<i>p</i>	D	0.92	0.87	0.71	0.81	0.77	0.93	0.88	0.97	0.844	0.91	0.87	0.99
		AH	0.94	0.88	0.72	0.82	0.90	0.94	0.91	0.98	0.99	0.96	0.93	0.99
		NP	0.83	0.78	0.66	0.69	0.68	0.81	0.72	0.94	0.84	0.83	0.74	0.97
	<i>s</i>	D	20e-4	22e-3	55e-3	54e-3	13e-4	20e-4	23e-4	17e-4	11e-4	15e-4	17e-4	9e-4
		AH	18e-4	19e-3	53e-3	54e-3	9e-4	18e-4	21e-4	13e-4	7e-4	12e-4	14e-4	4e-4
		NP	26e-4	11e-3	13e-3	49e-4	15e-4	26e-4	31e-4	43e-4	11e-4	19e-4	23e-4	41e-4

TAB. 5.3 – Table des valeurs des proportions d’estimations correctes selon le critère CC_2 , avec un intervalle de confiance $\hat{I}_{n,i}$ basé sur l’écart-type s pour différentes fréquences ω_0 , signaux g et SNR, avec un échantillonnage **Ech₁** et $\eta = 1$. Les échantillons de données sont de tailles $n = 100$, pour le premier tableau, et $n = 200$, pour le second. Le modèle d’échantillonnage **Ech₁** est le processus uniforme.

			Ech₂, n = 100											
			SNR = 1				SNR = 2				SNR = 3			
			<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄
$\omega_0 = 0.1$	<i>p</i>	D	0.90	0.64	0.59	0.22	0.75	0.89	0.82	0.43	0.67	0.82	0.69	0.54
		AH	0.90	0.63	0.60	0.23	0.89	0.90	0.90	0.53	0.94	0.91	0.86	0.64
		NP	0.83	0.58	0.39	0.22	0.71	0.79	0.4	0.55	0.72	0.68	0.35	0.64
	<i>s</i>	D	57e-4	90e-4	95e-4	11e-3	40e-4	58e-4	68e-4	80e-4	38e-4	47e-4	59e-4	71e-4
		AH	55e-4	94e-4	92e-4	11e-3	28e-4	56e-4	58e-4	72e-4	19e-4	37e-4	44e-4	62e-4
		NP	71e-4	88e-4	87e-4	88e-4	45e-4	67e-4	62e-4	89e-4	34e-4	52e-4	50e-4	91e-4
$\omega_0 = 0.8$	<i>p</i>	D	0.91	0.61	0.43	0.26	0.80	0.91	0.83	0.53	0.77	0.88	0.80	0.60
		AH	0.92	0.61	0.45	0.28	0.92	0.93	0.87	0.59	0.95	0.94	0.86	0.68
		NP	0.9	0.58	0.44	0.22	0.76	0.84	0.82	0.54	0.78	0.72	0.72	0.67
	<i>s</i>	D	61e-4	33e-3	41e-3	58e-4	35e-4	54e-4	67e-4	76e-4	30e-4	41e-4	50e-4	67e-4
		AH	52e-4	32e-3	37e-3	61e-3	26e-4	50e-4	62e-4	70e-4	17e-4	34e-4	43e-4	58e-4
		NP	63e-4	12e-3	17e-3	12e-3	38e-4	59e-4	71e-4	81e-4	29e-4	47e-4	50e-4	81e-4
$\omega_0 = 1.5$	<i>p</i>	D	0.91	0.58	0.40	0.29	0.81	0.90	0.83	0.57	0.77	0.88	0.80	0.63
		AH	0.93	0.57	0.42	0.30	0.92	0.92	0.86	0.63	0.95	0.93	0.86	0.71
		NP	0.92	0.56	0.4	0.26	0.77	0.82	0.78	0.56	0.78	0.73	0.64	0.68
	<i>s</i>	D	60e-4	72e-3	91e-3	12e-2	35e-4	74e-4	11e-3	77e-4	30e-4	41e-4	49e-4	63e-4
		AH	57e-4	70e-3	90e-3	11e-2	25e-4	51e-4	10e-3	65e-4	17e-4	34e-4	42e-4	54e-4
		NP	89e-4	34e-3	44e-3	38e-3	38e-4	67e-4	87e-4	83e-4	29e-4	48e-4	55e-4	81e-4

			Ech₂, n = 200											
			SNR = 1				SNR = 2				SNR = 3			
			<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄
$\omega_0 = 0.1$	<i>p</i>	D	0.93	0.91	0.82	0.73	0.77	0.94	0.84	0.95	0.83	0.91	0.79	0.97
		AH	0.91	0.89	0.83	0.77	0.87	0.94	0.90	0.97	0.97	0.95	0.89	0.99
		NP	0.79	0.74	0.42	0.69	0.71	0.60	0.45	0.95	0.83	0.48	0.18	0.98
	<i>s</i>	D	19e-4	34e-4	39e-4	46e-4	13e-4	19e-4	24e-4	22e-4	11e-4	15e-4	20e-4	15e-4
		AH	20e-4	37e-4	39e-4	41e-4	10e-4	19e-4	21e-4	15e-4	7e-4	13e-4	16e-4	8e-4
		NP	29e-4	45e-4	43e-4	48e-4	15e-4	28e-4	26e-4	45e-4	12e-4	22e-4	2e-4	45e-4
$\omega_0 = 0.8$	<i>p</i>	D	0.91	0.88	0.73	0.78	0.76	0.93	0.86	0.96	0.83	0.91	0.84	0.99
		AH	0.94	0.88	0.73	0.79	0.90	0.95	0.90	0.97	0.98	0.96	0.91	0.99
		NP	0.84	0.72	0.57	0.66	0.74	0.65	0.63	0.93	0.86	0.51	0.54	0.98
	<i>s</i>	D	20e-4	71e-4	14e-3	90e-4	13e-4	19e-4	24e-4	18e-4	11e-4	15e-4	18e-4	9e-4
		AH	18e-4	61e-4	15e-3	72e-4	9e-4	18e-4	21e-4	15e-4	7e-4	12e-4	15e-4	6e-4
		NP	26e-4	45e-4	93e-4	44e-4	14e-4	26e-4	29e-4	42e-4	11e-4	20e-4	20e-4	45e-4
$\omega_0 = 1.5$	<i>p</i>	D	0.91	0.84	0.70	0.80	0.77	0.91	0.86	0.96	0.84	0.90	0.83	0.98
		AH	0.94	0.85	0.70	0.81	0.89	0.94	0.87	0.98	0.98	0.96	0.89	0.99
		NP	0.83	1	0.66	0.69	0.74	1	0.73	0.93	0.86	1	0.71	0.97
	<i>s</i>	D	20e-4	22e-3	40e-3	38e-3	13e-4	20e-4	25e-4	20e-4	11e-4	15e-4	18e-4	14e-4
		AH	18e-4	22e-3	43e-3	40e-3	10e-4	18e-4	23e-4	14e-4	7e-4	12e-4	16e-4	8e-4
		NP	25e-4	68e-4	19e-3	23e-3	14e-4	26e-4	31e-4	45e-4	11e-4	19e-4	23e-4	43e-4

TAB. 5.4 – Table des valeurs des proportions d’estimations correctes selon le critère CC_2 , avec un intervalle de confiance $\hat{I}_{n,i}$ basé sur l’écart-type s pour différentes fréquences ω_0 , signaux g et SNR, avec un échantillonnage **Ech₂** et $\eta = 1$. Les échantillons de données sont de tailles $n = 100$, pour le premier tableau, et $n = 200$, pour le second. Le modèle d’échantillonnage **Ech₂** est le processus de Poisson.

		Ech₃, n = 100												
		SNR = 1				SNR = 2				SNR = 3				
		<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	
$\omega_0 = 0.1$	<i>p</i>	D	0.79	0.47	0.35	0.31	0.70	0.82	0.72	0.71	0.31	0.68	0.60	0.81
		AH	0.90	0.46	0.36	0.34	0.95	0.91	0.81	0.78	0.71	0.88	0.72	0.90
		NP	0.75	0.43	0.19	0.28	0.77	0.79	0.19	0.62	0.42	0.7	0.25	0.7
	<i>s</i>	D	20e-4	33e-4	38e-4	66e-4	14e-4	1e-49	21e-4	44e-4	13e-4	16e-4	19e-4	35e-4
		AH	16e-4	34e-4	40e-4	52e-4	8e-4	15e-4	19e-4	33e-4	6e-4	10e-4	15e-4	26e-4
		NP	22e-4	38e-4	40e-4	40e-4	13e-4	22e-4	22e-4	28e-4	11e-4	16e-4	18e-4	26e-4
$\omega_0 = 0.8$	<i>p</i>	D	0.84	0.44	0.29	0.37	0.76	0.82	0.74	0.84	0.34	0.73	0.65	0.93
		AH	0.90	0.44	0.28	0.40	0.95	0.90	0.80	0.89	0.73	0.88	0.80	0.97
		NP	0.79	0.42	0.3	0.31	0.84	0.79	0.7	0.73	0.47	0.68	0.67	0.81
	<i>s</i>	D	11e-3	38e-3	47e-3	11e-2	13e-4	65e-4	10e-3	45e-3	13e-4	16e-4	19e-4	26e-3
		AH	60e-4	33e-3	49e-3	11e-2	8e-4	15e-4	89e-4	26e-3	5e-4	10e-4	13e-4	17e-3
		NP	46e-4	13e-3	17e-3	23e-3	12e-4	33e-4	37e-4	70e-4	9e-4	15e-4	17e-4	64e-4
$\omega_0 = 1.5$	<i>p</i>	D	0.84	0.43	0.25	0.41	0.77	0.83	0.74	0.88	0.36	0.71	0.66	0.97
		AH	0.93	0.44	0.26	0.43	0.97	0.92	0.82	0.92	0.74	0.90	0.80	0.98
		NP	0.78	0.42	0.28	0.34	0.82	0.81	0.72	0.77	0.47	0.73	0.66	0.85
	<i>s</i>	D	28e-3	99e-4	12e-2	12e-2	13e-4	18e-3	34e-3	54e-3	12e-4	15e-4	11e-3	33e-3
		AH	19e-3	95e-3	13e-2	19e-2	7e-4	98e-4	34e-3	48e-3	5e-4	9e-4	58e-4	42e-4
		NP	23e-4	39e-3	51e-3	50e-3	12e-4	46e-4	11e-3	13e-3	9e-4	14e-4	17e-4	24e-4

		Ech₃, n = 200												
		SNR = 1				SNR = 2				SNR = 3				
		<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	
$\omega_0 = 0.1$	<i>p</i>	D	0.63	0.82	0.72	0.78	0.74	0.63	0.50	0.97	0.77	0.68	0.55	0.98
		AH	0.74	0.84	0.74	0.79	0.96	0.76	0.61	0.98	0.99	0.90	0.71	0.99
		NP	0.59	0.72	0.32	0.68	0.85	0.56	0.53	0.89	0.94	0.7	0.62	0.9
	<i>s</i>	D	6e-4	13e-4	17e-4	22e-4	5e-4	7e-4	8e-4	*	5e-4	6e-4	7e-4	*
		AH	5e-4	16e-4	16e-4	23e-4	2e-4	5e-4	6e-4	*	1e-4	3e-4	5e-4	*
		NP	7e-4	15e-4	17e-4	19e-4	4e-4	7e-4	8e-4	12e-4	2e-4	6e-4	6e-4	11e-4
$\omega_0 = 0.8$	<i>p</i>	D	0.61	0.79	0.62	0.83	0.71	0.63	0.53	0.99	0.74	0.70	0.61	1
		AH	0.76	0.83	0.63	0.84	0.97	0.76	0.63	0.99	0.99	0.91	0.80	1
		NP	0.56	0.73	0.58	0.71	0.87	0.57	0.55	0.94	0.95	0.73	0.67	0.97
	<i>s</i>	D	7e-4	12e-3	24e-3	48e-3	5e-4	7e-4	8e-4	*	5e-4	6e-4	6e-4	*
		AH	5e-4	11e-3	24e-3	47e-3	2e-4	5e-4	6e-4	*	1e-4	3e-4	4e-4	*
		NP	7e-4	18e-4	53e-4	11e-3	4e-4	7e-4	8e-4	10e-4	2e-4	5e-4	6e-4	10e-4
$\omega_0 = 1.5$	<i>p</i>	D	0.63	0.80	0.60	0.89	0.76	0.64	0.54	1	0.8	0.73	0.64	1
		AH	0.73	0.81	0.62	0.89	0.97	0.74	0.61	1	1	0.90	0.78	1
		NP	0.59	0.7	0.54	0.78	0.9	0.56	0.48	0.94	0.96	0.72	0.64	0.95
	<i>s</i>	D	6e-4	35e-3	59e-3	86e-3	5e-4	6e-4	7e-4	*	4e-4	5e-4	6e-4	*
		AH	5e-4	33e-3	59e-3	8e-2	1e-4	5e-4	7e-4	*	*	3e-4	5e-4	*
		NP	7e-4	15e-3	21e-3	77e-4	3e-4	7e-4	8e-4	10e-4	2e-4	5e-4	6e-4	10e-4

TAB. 5.5 – Table des valeurs des proportions d’estimations correctes selon le critère CC_2 , avec un intervalle de confiance $\hat{I}_{n,i}$ basé sur l’écart-type s pour différentes fréquences ω_0 , signaux g et SNR, avec un échantillonnage **Ech₃** et $\eta = 4.06$. Les échantillons de données sont de tailles $n = 100$, pour le premier tableau, et $n = 200$, pour le second. Le symbole *, veut dire que la valeur de s est inférieure à 10^{-4} . Le modèle d’échantillonnage **Ech₃** est le processus de mixture d’exponentielles.

			Ech₄, n = 100											
			SNR = 1				SNR = 2				SNR = 3			
			<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄
$\omega_0 = 0.1$	<i>p</i>	D	0.55	0.50	0.37	0.12	0.39	0.52	0.29	0.06	0.25	0.55	0.3	0.07
		AH	0.54	0.40	0.31	0.13	0.60	0.54	0.31	0.07	0.59	0.66	0.34	0.07
		NP	0.33	0.21	0.15	0.04	0.29	0.33	0.18	0.08	0.21	0.28	0.15	0.10
	<i>s</i>	D	11e-3	12e-3	14e-3	22e-3	11e-3	11e-3	13e-3	22e-3	11e-3	11e-3	13e-3	21e-3
		AH	11e-3	13e-3	13e-3	21e-3	90e-4	11e-4	12e-3	21e-3	75e-3	10e-3	12e-3	21e-3
		NP	10e-3	10e-3	10e-3	91e-4	99e-4	10e-3	11e-3	97e-4	97e-4	97e-4	10e-4	99e-4
$\omega_0 = 0.8$	<i>p</i>	D	0.57	0.49	0.38	0.24	0.44	0.60	0.49	0.17	0.34	0.67	0.58	0.18
		AH	0.78	0.50	0.38	0.27	0.84	0.76	0.58	0.21	0.84	0.91	0.77	0.25
		NP	0.69	0.35	0.24	0.09	0.72	0.69	0.53	0.18	0.58	0.70	0.50	0.22
	<i>s</i>	D	26e-3	47e-3	56e-3	1e-1	18e-3	23e-3	30e-3	58e-3	16e-3	19e-3	25e-3	43e-3
		AH	16e-3	44e-3	55e-3	98e-3	62e-4	16e-3	24e-3	52e-3	42e-4	91e-4	15e-3	41e-3
		NP	18e-3	28e-3	32e-3	28e-3	1e-2	16e-2	21e-3	21e-3	83e-4	13e-3	17e-3	20e-3
$\omega_0 = 1.5$	<i>p</i>	D	0.65	0.50	0.34	0.35	0.53	0.65	0.54	0.23	0.41	0.74	0.68	0.28
		AH	0.78	0.52	0.35	0.35	0.85	0.77	0.63	0.27	0.83	0.91	0.83	0.33
		NP	0.72	0.36	0.28	0.10	0.74	0.71	0.58	0.25	0.60	0.70	0.64	0.30
	<i>s</i>	D	30e-3	9e-2	12e-2	16e-2	19e-3	29e-3	45e-3	70e-3	15e-3	15e-3	24e-3	45e-3
		AH	16e-3	84e-3	11e-2	15e-2	65e-4	22e-3	35e-3	52e-3	42e-4	88e-4	13e-3	27e-3
		NP	20e-3	46e-3	57e-3	55e-3	92e-4	19e-3	25e-3	23e-3	71e-4	12e-3	16e-3	23e-3

			Ech₄, n = 200											
			SNR = 1				SNR = 2				SNR = 3			
			<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄
$\omega_0 = 0.1$	<i>p</i>	D	0.58	0.78	0.63	0.16	0.31	0.57	0.43	0.24	0.19	0.42	0.29	0.26
		AH	0.77	0.81	0.69	0.22	0.74	0.77	0.57	0.34	0.70	0.72	0.43	0.37
		NP	0.62	0.70	0.32	0.17	0.43	0.53	0.24	0.26	0.31	0.49	0.16	0.28
	<i>s</i>	D	66e-4	85e-4	97e-4	13e-3	61e-4	70e-4	88e-4	12e-3	60e-4	66e-4	88e-4	12e-3
		AH	50e-4	82e-4	87e-4	13e-3	27e-4	49e-4	70e-4	11e-3	19e-4	36e-4	66e-4	11e-3
		NP	61e-4	81e-4	84e-4	80e-4	47e-4	60e-4	72e-4	79e-4	43e-4	53e-4	70e-4	79e-4
$\omega_0 = 0.8$	<i>p</i>	D	0.67	0.80	0.64	0.36	0.40	0.65	0.60	0.54	0.27	0.51	0.46	0.59
		AH	0.85	0.86	0.68	0.39	0.82	0.86	0.73	0.62	0.78	0.83	0.67	0.67
		NP	0.69	0.71	0.59	0.31	0.52	0.59	0.50	0.53	0.41	0.55	0.52	0.57
	<i>s</i>	D	94e-4	20e-3	30e-3	53e-3	63e-4	65e-4	84e-4	18e-3	62e-4	54e-4	60e-4	21e-3
		AH	41e-4	16e-3	26e-3	40e-3	19e-4	38e-4	54e-4	12e-3	13e-4	26e-4	40e-4	86e-4
		NP	55e-4	10e-3	14e-3	96e-4	37e-4	56e-4	67e-4	83e-4	31e-4	45e-4	50e-4	82e-4
$\omega_0 = 1.5$	<i>p</i>	D	0.74	0.82	0.64	0.45	0.48	0.72	0.65	0.62	0.32	0.59	0.53	0.65
		AH	0.87	0.84	0.67	0.47	0.85	0.87	0.76	0.69	0.80	0.85	0.70	0.75
		NP	0.74	0.69	0.64	0.34	0.57	0.58	0.60	0.62	0.46	0.56	0.50	0.69
	<i>s</i>	D	76e-4	39e-3	69e-3	77e-3	42e-4	52e-4	15e-3	79e-4	39e-4	45e-4	52e-4	62e-4
		AH	37e-4	38e-3	66e-3	75e-3	18e-4	37e-4	50e-4	58e-4	12e-4	25e-4	36e-4	52e-4
		NP	52e-4	21e-3	24e-3	20e-3	35e-4	52e-4	64e-4	81e-4	26e-4	44e-4	55e-4	81e-4

TAB. 5.6 – Table des valeurs des proportions d’estimations correctes selon le critère CC_2 , avec un intervalle de confiance $\hat{I}_{n,i}$ basé sur l’écart-type s pour différentes fréquences ω_0 , signaux g et SNR, avec un échantillonnage **Ech₄** et $\eta = 0.610$. Les échantillons de données sont de tailles $n = 100$, pour le premier tableau, et $n = 200$, pour le second. Le modèle d’échantillonnage **Ech₄** est le processus de type Neyman-Scott.

Analyse des résultats pour l'échantillonnage Ech₂ avec le signal g_6

Le signal g_6 a été construit de manière à ce qu'il ne contienne plus l'harmonique de fréquence ω_0 , tout en étant périodique de période P_0 . Les histogrammes 5.2, pour $n = 100$ observations et 5.3, pour $n = 200$ observations, montrent que les méthodes de D et AH ne détectent pas la fréquence ω_0 , mais les harmoniques de fréquences supérieures. Tandis que la méthode NP est capable d'estimer la bonne fréquence.

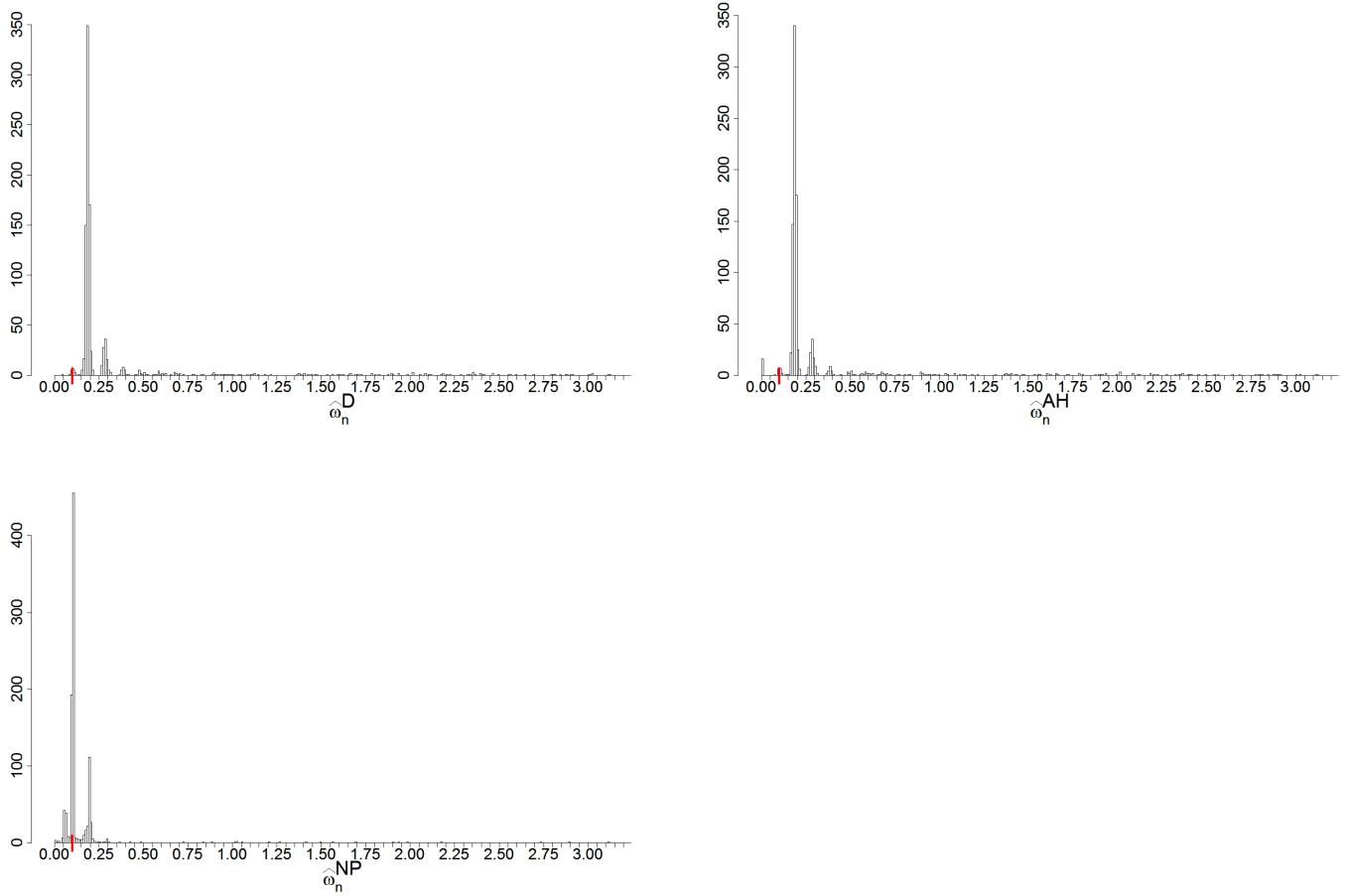


FIG. 5.2 – Histogrammes des estimations $\hat{\omega}_n^D$, $\hat{\omega}_n^{AH}$ et $\hat{\omega}_n^{NP}$, avec un échantillonnage Ech₂. La taille des échantillons vaut $n = 100$. La fréquence du signal g_6 vaut $\omega_0 = 0.1$ (barre verticale rouge.)

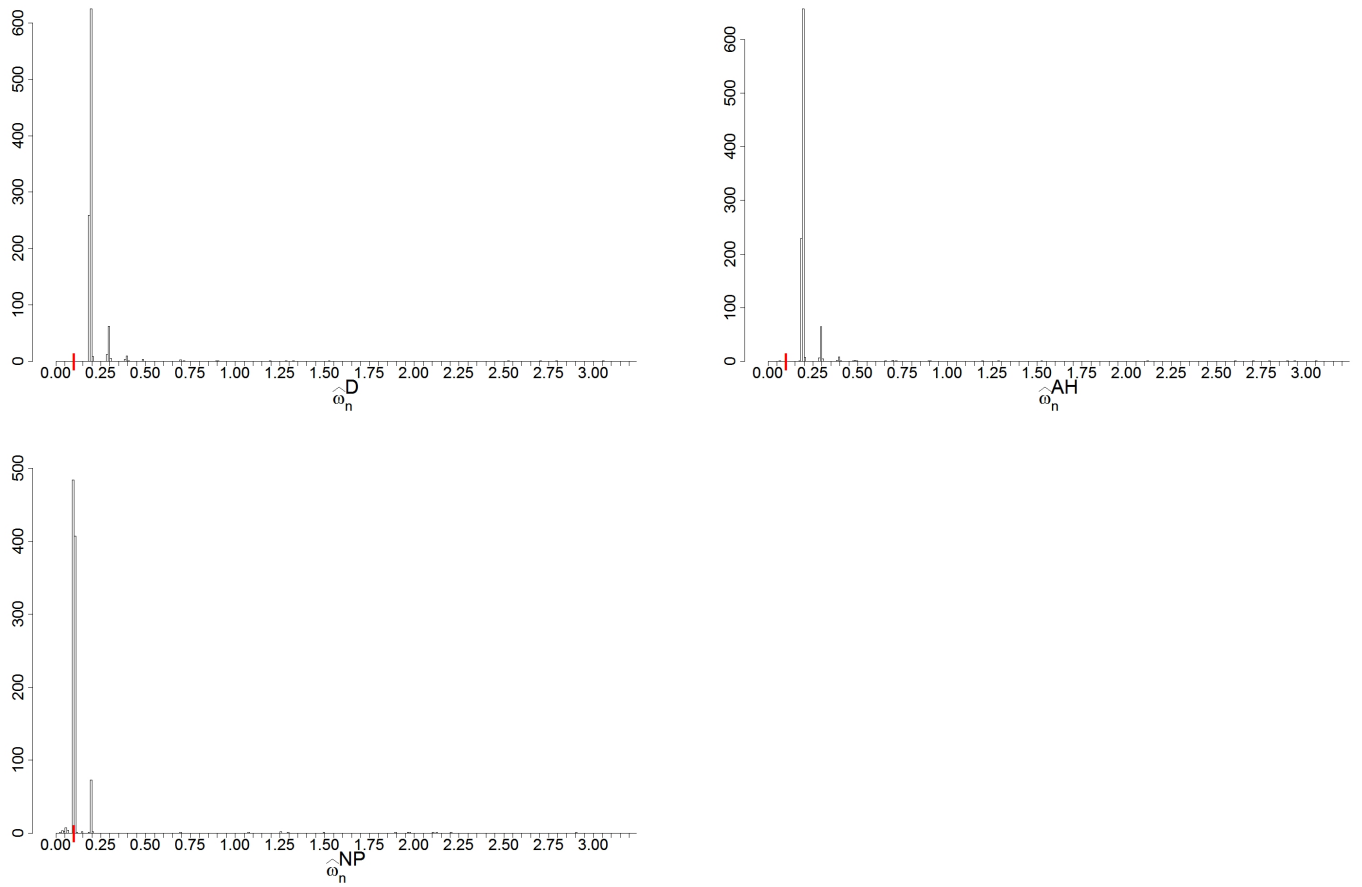


FIG. 5.3 – Histogrammes des estimations $\hat{\omega}_n^D$, $\hat{\omega}_n^{AH}$ et $\hat{\omega}_n^{NP}$, avec un échantillonnage Ech_2 . La taille des échantillons vaut $n = 200$. La fréquence du signal g_6 vaut $\omega_0 = 0.1$ (barre verticale rouge).

5.2.2 Modèles d'échantillonnage ne vérifiant pas la couverture de la période

Dans cette section, nous considérons quelques exemples d'échantillonnages ne vérifiant pas la condition de couverture de la période. La forme générale de ces temps aléatoires est le modèle (5.7).

$$T_l = l\delta + U_l, \quad (5.7)$$

où $l \in \mathbb{N}$, $\delta > 0$, $U_0 \sim \mathcal{U}(0, \beta P_0)$, $U_l = U \sim \mathcal{U}(-\alpha P_0, \beta P_0)$, $l > 0$, $0 < \alpha, \beta < \frac{\delta}{2}$. Dans le premier modèle, EchNU₁, on considère trois sous-modèles de (5.7), avec $\alpha = \beta$.

Type A $\delta = 3\frac{P_0}{8}$, $\alpha = \beta = \frac{1}{16}$,

Type B $\delta = 3\frac{P_0}{8}$, $\alpha = \beta = \frac{1}{32}$,

Type C $\delta = 3\frac{P_0}{8}$, $\alpha = \beta = \frac{1}{64}$.

L'échantillonnage de type A vérifie juste encore la condition de la couverture de la période. En effet, si l'on considère seulement l'espérance de chaque temps $\mathbb{E}[t_k] = k\delta$, après $k = 8$ observations, on a que $k\delta = 3P_0$. Ainsi, en considérant l'espérance des temps, les valeurs échantillonnées du signal g sont $g(0), g(\delta), g(2\delta), \dots, g(7\delta)$. Après un repliement en phase des observations, on obtient les mêmes valeurs $g(k\delta) = g(\frac{k}{8}\delta)$, aux temps $\frac{k}{8}\delta$, $k = 0, 1, \dots, 7$. L'intervalle entre deux temps successifs vaut $\frac{1}{8}$. Maintenant si l'on considère à nouveau les temps aléatoires, après un repliement en phase, on obtient des temps répartis sur les intervalles $[g(0), \frac{P_0}{16})$, $(g(k\delta) - \frac{P_0}{16}), g(k\delta) + \frac{P_0}{16})$, $k = 1, \dots, 6$ et $(g(P_0) - \frac{P_0}{16}), g(P_0))$, ce qui couvre tout l'intervalle $[0, P_0)$, excepté les temps $\frac{P_0}{16} + k\frac{P_0}{8}$, $k = 0, \dots, 7$.

Pour les échantillonnages de type B et C, nous gardons la même espérance des temps $\delta = 3\frac{P_0}{8}$, mais on considère un intervalle, autour des valeurs moyennes du signal observé, dont sa longueur $\frac{1}{8}$ est divisée par 2 et 4 respectivement. Ainsi la condition de couverture de la période n'est plus vérifiée. La figure 5.4 montre un exemple des signaux g_2 et g_3 échantillonnés avec les types d'échantillonnages B et C. Les résultats des simulations sont affichés dans les tableaux 5.7 à 5.9.

Pour le deuxième modèle d'échantillonnage EchNU₂, on considère les sous-modèles de (5.7), avec $\alpha = 0$.

Type A $\delta = P_0$, $\beta = \frac{1}{2}$, $\alpha = 0$,

Type B $\delta = P_0$, $\beta = \frac{1}{4}$, $\alpha = 0$,

Type C $\delta = P_0$, $\beta = \frac{1}{8}$, $\alpha = 0$.

Ces modèles d'échantillonnages concentrent les observations sur une même portion du signal. Les type A, B et C ne permettent l'observation du signal que sur la première moitié, premier quart et premier huitième respectivement. La figure 5.5 montre un exemple des signaux g_2 et g_3 échantillonnés avec les types d'échantillonnages B et C. Les résultats des simulations sont affichés dans les tableaux 5.10 à 5.12. La fréquence considérée est $\omega_0 = 0.1$.

Analyse des résultats pour l'échantillonnage EchNU₁ (tableaux 5.7 à 5.9)

En considérant la formule de la variance asymptotique de τ_{NP}^2 ((4.132)), théorème 12), avec l'échantillonnage EchNU₁, $L = 8$, $\delta = \frac{3P_0}{8}$. Pour l'échantillonnage de type A, tableau 5.7, $\beta = \frac{1}{16}$. Ce modèle d'échantillonnage vérifie juste encore la condition de la couverture de la période. Les valeurs $s_{\text{AH},1}$ et $s_{\text{AH},2}$ peuvent donc être comparées aux valeurs τ_{AH} du tableau 5.2. Il faut encore diviser τ_{AH} par

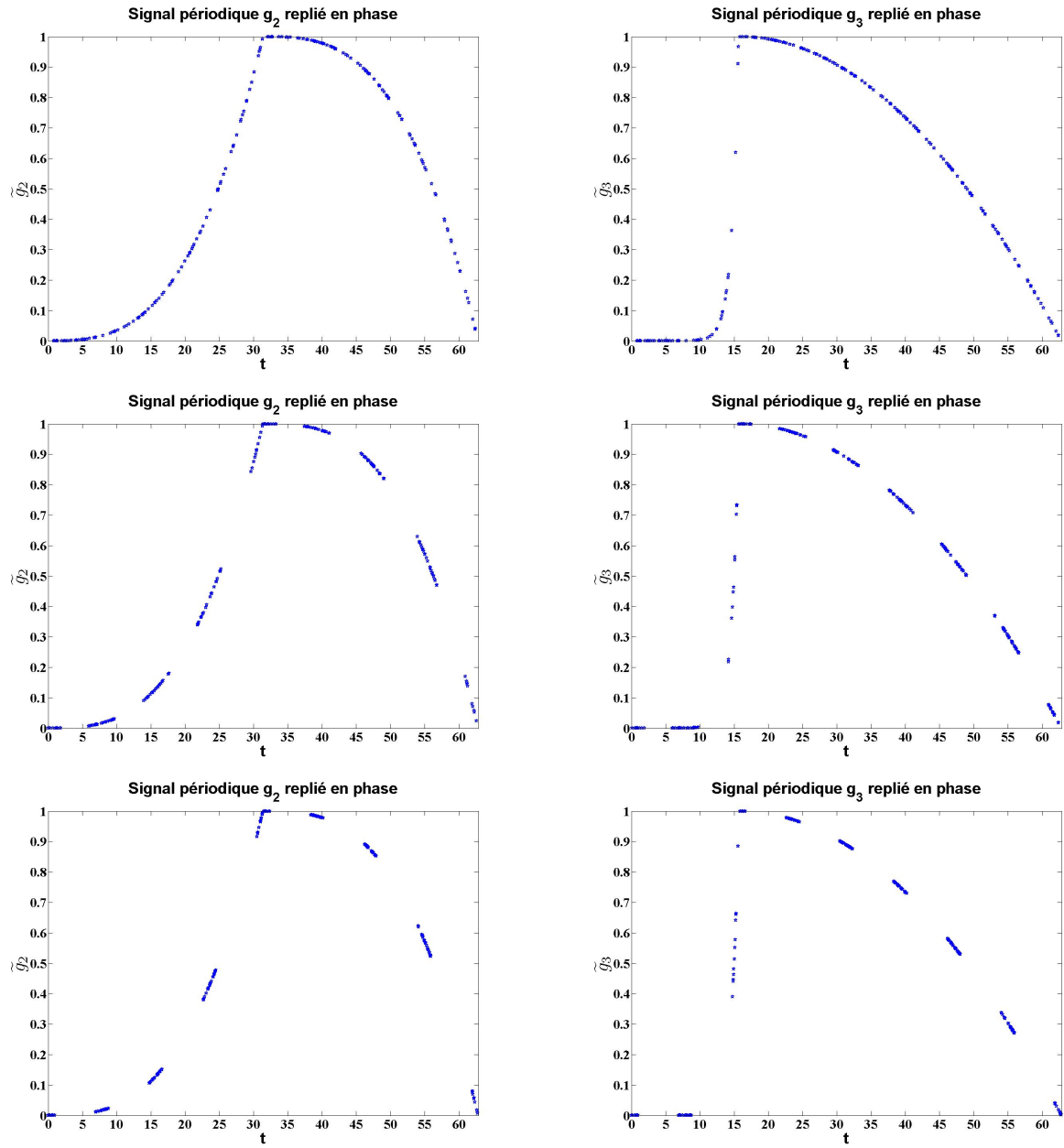


FIG. 5.4 – Observations des signaux g_2 et g_3 repliés en phase, avec un modèle d'échantillonnage EchNU₁ de type A de type B (première ligne) de type B (deuxième ligne) et C (dernière ligne).

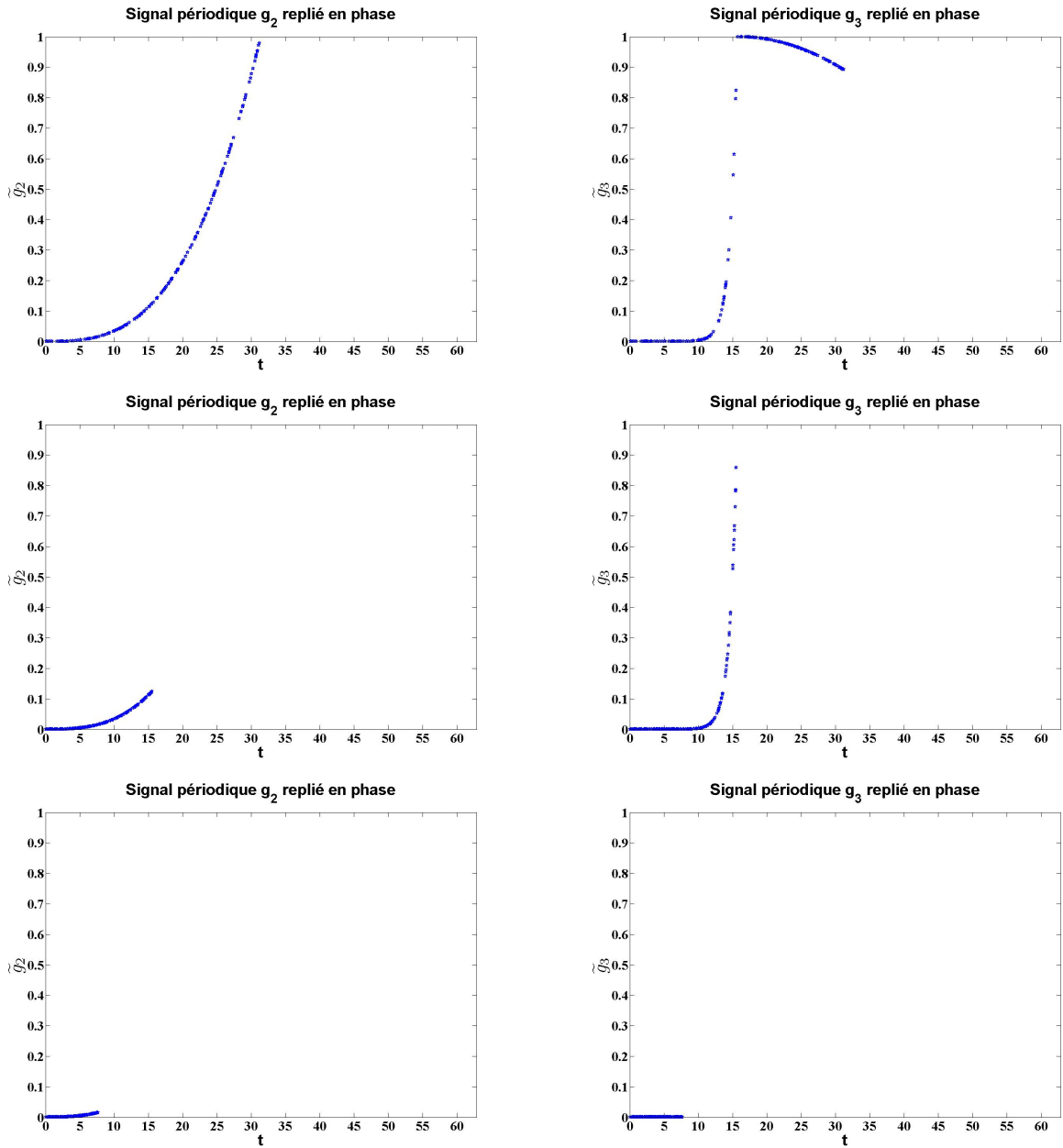


FIG. 5.5 – Observations des signaux g_2 et g_3 repliés en phase, avec des modèles d'échantillonnages EchNU₂ de type A (première ligne), B (deuxième ligne) et C (dernière ligne).

$\eta = \delta = \frac{3}{8}P_0$. Etant données les mauvaises performances de la méthode NP, on ne peut pas effectuer de comparaison entre s_{NP} et τ_{NP} pour $\omega_0 = 0.1$. Pour l'échantillonnage de type B, tableau 5.8, $\beta = \frac{1}{32}$. On obtient que la méthode AH est supérieure aux autres méthodes, pour $n = 100$ et $n = 200$ données. La comparaison des valeurs de s_{AH} et S_D a montré que lorsque s_{AH} est plus petite que S_D et si ce n'est pas le cas, la différence est insignifiante. Effectuons une comparaison de certaines valeurs de s avec la théorie asymptotique du théorème 12 pour le cas d'un échantillonnage ne vérifiant pas la couverture de la période. Considérons le tableau 5.8, lorsque $SNR = 1$, avec signal g_1 de fréquence $\omega_0 = 0.1$ et $n = 100$, $s_{NP} = 0.0018$. L'intégrale sur les intervalles couverts par les observations vaut $Int_{g_1} \simeq 0.1570$. Ainsi la racine de la variance asymptotique vaut

$$\sqrt{2\pi L \frac{1}{n^3} \omega_0 \frac{1}{\delta^2} 2\beta \frac{12}{Int_{g_1}}} \simeq 0.000207,$$

qui est en fait égale

$$\sqrt{\frac{24}{n^3} \frac{1}{\delta^2} \sigma^2} \simeq 0.000207.$$

Dans le tableau 5.8, lorsque $n = 200$, $SNR = 1$, avec signal g_1 de fréquence $\omega_0 = 0.1$, $s_{NP} = 0.0017$. L'intégrale sur les intervalles couverts par les observations vaut toujours $Int_{g_1} \simeq 0.1570$. Ainsi la racine de la variance asymptotique vaut

$$\sqrt{2\pi L \frac{1}{n^3} 2\beta \omega_0 \frac{1}{\delta^2} \frac{12}{Int_{g_1}}} \simeq 0.000073,$$

qui est en fait égale à

$$\sqrt{\frac{24}{n^3} \frac{1}{\delta^2} \sigma^2} \simeq 0.000073.$$

A présent considérons le signal g_2 avec la fréquence $\omega_0 = 0.1$. Lorsque $SNR = 1$ et $n = 100$, $s_{NP} = 0.0036$. L'intégrale sur les intervalles couverts par les observations vaut $Int_{g_2} \simeq 0.059$. Ainsi la racine de la variance asymptotique vaut

$$\sqrt{2\pi L \frac{1}{n^3} 2\beta \omega_0 \frac{1}{\delta^2} \frac{12}{Int_{g_1}}} \simeq 0.000334,$$

qui est en fait égale

$$\sqrt{\frac{2\pi 12}{n^3} \frac{1}{\int_0^{P_0} g_1'(u)^2 du} \frac{1}{\delta}} \simeq 0.000334.$$

Avec $n = 200$, on a obtenu que $s_{NP} = 0.0034$. L'intégrale sur les intervalles couverts par les observations vaut $Int_{g_2} = 0.059$. Ainsi la racine de la variance asymptotique vaut

$$\sqrt{2\pi L 2\beta \frac{1}{n^3} \omega_0 \frac{1}{\delta^2} \frac{12}{Int_{g_2}}} \simeq 0.000119,$$

qui est en fait égale à

$$\sqrt{\frac{2\pi 12}{n^3} \frac{1}{\delta^2} \frac{1}{\int_0^{P_0} g'_2(u)^2 du}} \simeq 0.000119.$$

Ces relations montrent que le modèle d'échantillonnage EchNU₁ de type A (vérifiant la condition de couverture de la période), B et C (ne la vérifiant pas), donnent la même variance asymptotique τ_{NP}^2 pour les signaux g_1 et g_2 . A la suite de la démonstration du théorème 12, à la section 4.3, nous avons montré que la variance asymptotique τ_{NP} , pour le signal g_1 , avec le modèle d'échantillonnage EchNU₁ et $\beta \rightarrow 0$, tend vers la même variance asymptotique. On observe le même phénomène pour le signal g_2 . Cela ne se vérifie pas pour le signal g_3 et g_4 . En effet, la composante sinusoidale de fréquence ω_0 explique presque complètement le signal g_2 , alors que pour le signal g_3 la composante de la fréquence fondamentale ω_0 ne suffit pas. Cela est encore plus marqué pour le signal g_4 , où la composante sinusoidale de fréquence $2\omega_0$ est la plus présente. Pour effectuer ces comparaisons nous avons dû considérer les estimations $\hat{\omega}_n^{NP}$ dans un voisinage assez proche de la fréquence ω_0 . Cela pour ne pas prendre en compte les mauvaises estimations de la fréquence, dues au fait que l'estimation de la fonction de régression n'est pas toujours assez précise. Cependant on voit que nos valeurs de s sont assez éloignées de la variance asymptotique. Dans la dernière section (5.5), nous présentons des simulations effectuées avec la fonction de régression connue, ce qui nous permettra de mieux vérifier nos formules en les comparant aux valeurs de s .

Analyse des résultats pour l'échantillonnage EchNU₂ (tableau 5.10)

Pour ce modèle d'échantillonnage, pour la comparaison des méthodes, nous considérons un intervalle de confiance $I_{n,i}$, $i = 1, \dots, N$, N le nombre de simulations, comme précédemment, mais basé sur la variance asymptotique (4.132), où $L = 1$, et l'intégrale $Int_g = \int_0^{\beta P_0} g'(u)^2 du$, ce qui correspond à la partie du signal étant couvert par l'échantillonnage EchNU₂. Si $\omega_0 \in I_{n,i}$, alors on incrémente de 1 le nombre d'estimations considérées comme bonnes. Ce modèle d'échantillonnage permet l'observation du signal que sur une partie du signal. L'échantillonnage de type A (tableau 5.10) permet l'observation du signal que sur sa première moitié. Les résultats obtenus par la méthode AH sont encore bons pour les signaux g_1 , g_2 et g_4 . Le manque d'observations du signal g_3 diminue fortement l'efficacité de la méthode AH. D'ailleurs pour des valeurs du SNR = 2, 3, la performance de la méthode NP est supérieure à la performance de la méthode AH. L'échantillonnage de type B (tableau 5.11) permet l'observation du signal que sur son premier quart. La performance de la méthode NP est clairement supérieure aux deux autres, excepté pour le signal g_1 , où la méthode D a une meilleure performance pour SNR = 2, 3. L'échantillonnage de type B permet l'observation que de la partie exponentielle du signal, tandis que l'échantillonnage de type A permet aussi d'observer une petite partie de la partie quadratique du signal. L'observation de cette partie quadratique avec l'ajout d'erreurs induit en erreur la méthode AH. L'échantillonnage de type C (tableau 5.12) permet l'observation du signal que sur son premier huitième. La méthode NP prend largement le dessus. Cependant les performances restent basses. Pour ce type d'échantillonnage, nous n'avons pas effectué de considération asymptotique. Cependant il doit être possible d'en dériver une, à partir de la formule (4.132) du théorème 12.

		EchNU₁, type A, n = 100											
		SNR = 1				SNR = 2				SNR = 3			
		<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄
<i>p</i>	D	0.98	0.61	0.54	0.27	1	0.98	0.96	0.83	1	0.99	1	0.92
	AH	0.97	0.49	0.38	0.58	1	0.98	0.96	0.98	1	0.99	0.99	0.99
	NP	0.99	0.34	0.33	0.58	1	0.97	0.95	0.99	1	1	1	1
<i>s</i>	D	1e-4	18e-4	22e-4	44e-4	*	1e-4	2e-4	2e-4	*	*	*	1e-4
	AH	1e-4	1e-3	14e-4	25e-4	*	1e-4	2e-4	13e-4	*	*	*	*
	NP	*	14e-4	13e-4	86e-04	*	1e-4	2e-4	*	*	*	*	*

		EchNU₁, type A, n = 200											
		SNR = 1				SNR = 2				SNR = 3			
		<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄
<i>p</i>	D	1	0.97	0.93	0.77	1	1	1	0.98	1	1	1	0.99
	AH	1	0.94	0.86	0.92	1	1	1	0.99	1	1	1	1
	NP	1	0.84	0.83	0.93	1	1	1	1	1	1	1	1
<i>s</i>	D	*	5e-4	8e-4	1e-4	*	*	*	*	*	*	*	*
	AH	*	3e-4	5e-4	13e-4	*	*	*	*	*	*	*	*
	NP	*	8e-4	6e-4	*	*	*	*	*	*	*	*	*

TAB. 5.7 – Table des valeurs des proportions p d’estimations correctes selon le critère CC_2 , avec un intervalle de confiance $I_{n,i}$, basé sur l’écart-type τ , pour $\omega_0 = 0.1$. L’échantillon de données de taille $n = 100$, pour le premier tableau, et $n = 200$, pour le second tableau. Le modèle d’échantillonnage est le modèle **EchNU₁** de type A. Le symbole * veut dire que la valeur de s est inférieure à 10^{-4} .

		EchNU₁, type B, n = 100											
		SNR = 1				SNR = 2				SNR = 3			
		<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄
<i>p</i>	D	0.97	0.49	0.37	0.09	1	0.98	0.96	0.41	1	0.99	0.99	0.58
	AH	0.97	0.49	0.36	0.26	1	0.98	0.96	0.45	1	0.99	0.99	0.58
	NP	0.91	0.32	0.30	0.67	1	0.94	0.93	0.99	1	1	0.99	1
<i>s</i>	D	1e-4	12e-4	13e-4	12e-3	*	1e-4	1e-4	28e-4	*	*	*	1e-4
	AH	2e-4	11e-4	12e-4	34e-4	*	1e-4	1e-4	*	*	*	*	*
	NP	2e-4	13e-4	15e-4	59e-4	*	2e-4	2e-4	*	*	*	1e-4	*

		EchNU₁, type B, n = 200											
		SNR = 1				SNR = 2				SNR = 3			
		<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄
<i>p</i>	D	1	0.92	0.82	0.41	1	1	0.99	0.45	1	1	1	0.60
	AH	1	0.92	0.82	0.37	1	1	0.99	0.79	1	1	1	0.90
	NP	1	0.86	0.82	0.90	1	1	1	1	1	1	1	1
<i>s</i>	D	*	0.0001	0.0004	0.0053	*	*	*	*	*	*	*	*
	AH	*	0.0001	0.0001	*	*	*	*	*	*	*	*	*
	NP	*	11e-4	8e-4	38e-4	*	*	*	*	*	*	*	*

TAB. 5.8 – Table des valeurs des proportions p d’estimations correctes selon le critère CC_2 , avec un intervalle de confiance $I_{n,i}$, basé sur l’écart-type τ , pour $\omega_0 = 0.1$. L’échantillon de données de taille $n = 100$, pour le premier tableau, et $n = 200$, pour le second tableau. Le modèle d’échantillonnage est le modèle **EchNU₁** de type B. Le symbole * veut dire que la valeur de s est inférieure à 10^{-4} .

		EchNU₁, type C, n = 100											
		SNR = 1				SNR = 2				SNR = 3			
		<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄
<i>p</i>	D	0.83	0.33	0.26	0.13	0.97	0.84	0.78	0.08	0.98	0.95	0.89	0.03
	AH	0.86	0.34	0.27	0.05	0.99	0.88	0.8	0.21	1	0.98	0.94	0.3
	NP	0.7	0.18	0.16	0.77	0.97	0.73	0.60	0.99	1	0.94	0.77	1
<i>s</i>	D	2e-4	17e-4	16e-4	15e-3	*	1e-4	2e-4	54e-4	*	*	*	26e-4
	AH	2e-4	16e-4	15e-4	*	*	1e-4	2e-4	*	*	*	*	*
	NP	2e-4	26e-4	24e-4	*	*	2e-4	2e-4	*	*	*	*	*

		EchNU₁, type C, n = 200											
		SNR = 1				SNR = 2				SNR = 3			
		<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄
<i>p</i>	D	0.95	0.74	0.61	0.12	1	0.97	0.91	0.03	1	0.99	0.96	0.01
	AH	0.97	0.77	0.63	0.20	1	0.98	0.92	0.43	1	0.99	0.99	0.51
	NP	0.7	0.18	0.16	0.77	0.97	0.73	0.60	0.99	1	0.94	0.77	1
<i>s</i>	D	*	4e-4	4e-4	81e-4	*	*	*	16e-4	*	*	*	*
	AH	*	21e-4	4e-4	*	*	*	*	*	*	*	*	*
	NP	2e-4	26e-4	24e-4	*	*	2e-4	2e-4	*	*	*	*	*

TAB. 5.9 – Table des valeurs des proportions p d'estimations correctes selon le critère CC_2 , avec un intervalle de confiance $I_{n,i}$, basé sur l'écart-type τ , pour $\omega_0 = 0.1$. L'échantillon de données de taille $n = 100$, pour le premier tableau, et $n = 200$, pour le second tableau. Le modèle d'échantillonnage est le modèle **EchNU₁** de type C. Le symbole * veut dire que la valeur de s est inférieure à 10^{-4} .

		EchNU₂, type A, n = 100											
		SNR = 1				SNR = 2				SNR = 3			
		<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄
<i>p</i>	D	0.88	0.30	0.37	0.08	0.99	0.96	0.42	0.14	1	1	0.39	0.15
	AH	0.60	0.19	0.30	0.02	0.73	0.78	0.36	0	0.77	0.92	0.34	0
	NP	0.92	0.23	0	0.51	1	0.90	0	0.82	1	1	0	0.90

		EchNU₂, type A, n = 200											
		SNR = 1				SNR = 2				SNR = 3			
		<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄
<i>p</i>	D	0.97	0.68	0.43	0.13	1	1	0.38	0.14	1	1	0.33	0.12
	AH	0.73	0.46	0.34	0	0.86	0.92	0.34	0	0.91	0.97	0.32	0
	NP	1	0.58	0	0.75	1	1	0	0.91	1	1	0	0.97

TAB. 5.10 – Table des valeurs des proportions p d'estimations correctes selon le critère CC_2 , avec intervalle $J_{n,i}$, pour $\omega_0 = 0.1$. L'échantillon de données de taille $n = 100$, pour le premier tableau, et $n = 200$, pour le second tableau. Le modèle d'échantillonnage est le modèle **EchNU₂** de type A.

		EchNU₂, type B, n = 100											
		SNR = 1				SNR = 2				SNR = 3			
		<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄
<i>p</i>	D	0.13	0.10	0.13	0	0.45	0.11	0.31	0	0.65	0.12	0.58	0
	AH	0.09	0.11	0.01	0	0.03	0.11	0.08	0	0.01	0.11	0.05	0
	NP	0.06	0.13	0	0.44	0.05	0.1	0	0.54	0.07	0.09	0	0.58

		EchNU₂, type B, n = 200											
		SNR = 1				SNR = 2				SNR = 3			
		<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄
<i>p</i>	D	0.236	0.109	0.183	0	0.631	0.112	0.606	0	0.713	0.121	0.896	0
	AH	0.067	0.102	0.091	0	0.002	0.102	0.036	0	0	0.096	0.014	0
	NP	0.05	0.15	0.03	0.25	0.03	0.12	0.25	0.3	0.03	0.12	0.29	0.31

TAB. 5.11 – Table des valeurs des proportions p d’estimations correctes selon le critère CC_2 , avec un intervalle $J_{n,i}$, pour $\omega_0 = 0.1$. L’échantillon de données de taille $n = 100$, pour le premier tableau, et $n = 200$, pour le second tableau. Le modèle d’échantillonnage est le modèle **EchNU₂** de type B.

		EchNU₂, type C, n = 100											
		SNR = 1				SNR = 2				SNR = 3			
		<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄
<i>p</i>	D	0.11	0.17	0.17	0.82	0.15	0.17	0.17	0.97	0.23	0.17	0.17	0.99
	AH	0.11	0.16	0.16	0.01	0.11	0.16	0.16	0	0.08	0.16	0.16	0
	NP	0.11	0.18	0	0.37	0.08	0.15	0	0.37	0.06	0.14	0	0.35

		EchNU₂, type C, n = 200											
		SNR = 1				SNR = 2				SNR = 3			
		<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄	<i>g</i> ₁	<i>g</i> ₂	<i>g</i> ₃	<i>g</i> ₄
<i>p</i>	D	0.111	0.161	0.161	0.951	0.204	0.162	0.161	0.994	0.349	0.163	0.161	1
	AH	0.101	0.167	0.166	0.001	0.086	0.165	0.166	0	0.046	0.165	0.166	0
	NP	0.04	0.14	0	0.2	0.04	0.11	0.01	0.21	0.03	0.12	0	0.17

TAB. 5.12 – Table des valeurs des proportions p d’estimations correctes selon le critère CC_2 , avec un intervalle $J_{n,i}$, pour $\omega_0 = 0.1$. L’échantillon de données de taille $n = 100$, pour le premier tableau, et $n = 200$, pour le second tableau. Le modèle d’échantillonnage est le modèle **EchNU₂** de type C.

5.3 Simulation de données avec erreurs gaussiennes corrélées

Dans cette section, nous testons la validité de la formule de la variance asymptotique (4.110) du théorème 11, dans le cas où les erreurs suivent un processus Gaussien corrélé. Pour tester la formule, nous considérons les fonctions g_1 , g_2 et g_3 avec fréquence $\omega_0 = 0.1$ et de rapport $\frac{R}{\sigma} = 2$, sur un échantillonnage de Poisson avec paramètre $\lambda = 1$. Les erreurs corrélées sont générées par processus stochastique autorégressif continu CAR(1) de premier ordre, défini au paragraphe 4.4.2. Pour la simulations de ces erreurs, nous utilisons la procédure récursive, proposée dans [20],

$$\varepsilon(t_n) = \exp(-\phi(t_n - t_{n-1})) \varepsilon(t_{n-1}) + \sigma (1 - \exp(-2\phi(t_n - t_{n-1})))^{\frac{1}{2}} \varepsilon^*(t_n). \quad (5.8)$$

où $\varepsilon^*(t)$ est un processus stochastique continu Gaussien indépendant. La densité spectrale $f_{sp}(\omega)$ d'un processus CAR(1) stationnaire, de paramètre $\phi > 0$, vaut

$$f_{sp}(\omega) = \frac{1}{\omega^2 + \phi^2}. \quad (5.9)$$

Nous considérons donc la méthode NP avec les fonctions g_i , $i = 1, 2, 3$ comme fonctions de régression connues. Dans ce cas, la formule de la variance asymptotique (4.110) devient

$$\tau^2 = 2\pi \frac{3}{n^3} \omega_0 \frac{\int_{-\infty}^{\infty} |\mathcal{TF}\{g'\}(\omega)|^2 f_{sp}(\frac{\omega}{P_0}) d\omega}{\left(\int_0^{P_0} g'(u)^2 du\right)^2}. \quad (5.10)$$

En effet, la fonction g n'étant plus un paramètre inconnu, il ne reste plus que la période P comme paramètre inconnu. Les calculs de la preuve du théorème 11 se simplifient. Pour un calcul formel, voir l'appendix 7. Nous calculons les statistiques s_{NP} des estimateurs $\hat{\omega}_n^{NP}$ de la fréquence ω_0 des fonctions g_i et les comparons avec la formule asymptotique. Maintenant, effectuons les calculs pour la formule de la variance asymptotique avec les fonctions g_1 , g_2 et g_3 . Pour le signal g_1 , selon la formule asymptotique (7.2) du théorème 13, avec $\omega_0 = 0.1$, $R = 2$ et

$$f_{sp}(\omega_0) = \begin{cases} 3.846 & \text{si } \phi = 0.5, \\ 1.219 & \text{si } \phi = 0.9. \end{cases}$$

On voit qu'un processus CAR(1) avec $0 < \phi < 1$ et une fréquence $\omega_0 = 0.1$, l'écart-type asymptotique τ est péjoré par rapport au cas où les erreurs sont iid. Pour $\phi = 0.9$, on obtient

$$\begin{aligned} n^{-\frac{3}{2}} \tau &= \sqrt{\frac{6}{R^2} f_{sp}(\omega_0)} \\ &\simeq \begin{cases} 0.00135 & \text{si } n = 100, \\ 0.00047 & \text{si } n = 200, \\ 0.00016 & \text{si } n = 400. \end{cases} \end{aligned}$$

Pour le signal g_2 , l'intégrale $\int_0^{P_0} g_2'(u)^2 du$ vaut

$$\int_0^{P_0} g_2'(u)^2 du \simeq 0.113R^2 = 0.4549 \text{ si } \omega_0 = 0.1 \text{ et } R = 2.$$

La transformée de Fourier $\mathcal{TF}\{g_2'\}(\omega)$ de g_2' s'écrit

$$\mathcal{TF}\{g'_2\}(\omega) = R(\mathcal{TF}\{g'_2\}_1(\omega) + \mathcal{TF}\{g'_2\}_2(\omega)),$$

où

$$\begin{aligned} \mathcal{TF}\{g'_2\}_1(\omega) &= \frac{1}{C} \left[3a \left(-\frac{1}{i\omega} \exp\left(-i\omega \frac{P_0}{2}\right) \left(\frac{P_0}{2}\right)^2 + \frac{2}{\omega^2} \exp\left(-i\omega \frac{P_0}{2}\right) \frac{P_0}{2} + \frac{2}{i\omega^3} \left(\exp\left(-i\omega \frac{P_0}{2}\right) - 1\right) \right) \right. \\ &\quad + 2b \left(-\frac{1}{i\omega} \exp\left(-i\omega \frac{P_0}{2}\right) \frac{P_0}{2} + \frac{1}{\omega^2} \left(\exp\left(-i\omega \frac{P_0}{2}\right) - 1\right) \right) \\ &\quad \left. - \frac{c}{i\omega} \left(\exp\left(-i\omega \frac{P_0}{2}\right) - 1\right) \right], \end{aligned}$$

avec $C = a\left(\frac{P_0}{2}\right)^3 + b\left(\frac{P_0}{2}\right)^2 + c\left(\frac{P_0}{2}\right)$. Et

$$\mathcal{TF}\{g'_2\}_2(\omega) = -\mathcal{TF}\{g'_2\}_1(\omega) \exp\left(-i\omega \frac{P_0}{2}\right),$$

et

$$\int_{-\infty}^{+\infty} \mathcal{TF}\{g'_2\}(\omega) f_{sp}\left(\frac{\omega}{P_0}\right) d\omega \simeq 2.853.$$

Par la formule (5.10),

$$n^{-\frac{3}{2}}\tau \simeq \begin{cases} 0.00225 & \text{si } n = 100, \\ 0.00079 & \text{si } n = 200, \\ 0.00028 & \text{si } n = 400. \end{cases}$$

Pour le signal g_3 , l'intégrale $\int_0^{P_0} g'_3(u)^2 du$ vaut

$$\int_0^{P_0} g'_3(u)^2 du \simeq 0.528R^2 = 2.113 \text{ si } \omega_0 = 0.1 \text{ et } R = 2$$

La transformée de Fourier $\mathcal{TF}\{g'_3\}(\omega)$ de g'_3 s'écrit

$$\begin{aligned} \mathcal{TF}\{g'_3\}(\omega) &= \exp\left(-\frac{P_0}{4}\right) \left(\frac{1}{1-i\omega} \exp\left(\left(1-i\omega\right)\frac{P_0}{4}\right) - 1 \right) \\ &\quad - \frac{8}{3P_0} \left(\frac{1}{\omega^2} \left(\exp(-i\omega P_0) - \exp\left(-i\omega \frac{P_0}{4}\right) \right) + \frac{1}{\omega} \left(-P_0 \exp(-i\omega P_0) + \frac{P_0}{4} \exp\left(-i\omega \frac{P_0}{4}\right) \right) \right) \\ &\quad - \frac{P_0}{4} \frac{i}{\omega} \left(\exp(-i\omega P_0) - \exp\left(-i\omega \frac{P_0}{4}\right) \right). \end{aligned}$$

Et

$$\int_{-\infty}^{+\infty} \mathcal{TF}\{g'_3\}(\omega) f_{sp}\left(\frac{\omega}{P_0}\right) d\omega \simeq 13.078,$$

Par la formule (5.10),

$$n^{-\frac{3}{2}}\tau \simeq \begin{cases} 0.00104 & \text{si } n = 100, \\ 0.00036 & \text{si } n = 200, \\ 0.00013 & \text{si } n = 400. \end{cases}$$

Les comparaisons des valeurs de la statistique s_{NP} ainsi que l'écart type asymptotique $n^{-\frac{3}{2}}\tau$, pour les fonctions g_i , $i = 1, 2, 3$, de fréquence $\omega_0 = 0.1$, sont présentés dans le tableau 5.13. Globalement, on remarque que la formule (4.110) de la variance asymptotique donne des valeurs assez proches de s . Nous comparons aussi les valeurs de s_{NP} avec $n^{-\frac{3}{2}}\tau$, avec des pour les fonctions g_i , $i = 1, 2, 3$. Le tableau 5.14 montre la comparaison des valeurs de s_{NP} avec $n^{-\frac{3}{2}}\tau$, lorsque les erreurs suivent un processus CAR(1) avec $\phi = 0.9$. Nous considérons à nouveau les trois fonctions g_1 , g_2 et g_3 , avec fréquences $\omega_0 \in \{0.1, 1.53\}$. Ces résultats montrent que les hautes fréquences sont en général beaucoup mieux estimées que le basses fréquences, lorsque les erreurs sont de type CAR(1) avec $\phi = 0.9$ et le signal est g_1 et g_2 . En effet, pour le signal g_1 , la fonction de densité spectrale pour $\omega = \omega_0$ vaut

$$f_{sp}(\omega_0) \simeq \begin{cases} 1.219 & \text{si } \omega_0 = 0.1, \\ 0.689 & \text{si } \omega_0 = 0.8, \\ 0.326 & \text{si } \omega_0 = 1.5. \end{cases}$$

Pour le signal g_3 , Le tableau montre que l'écart-type asymptotique est plus inférieur lorsque $\omega_0 = 0.1$. Cependant, pour des fréquences beaucoup plus grandes, par exemple $\omega_0 = 5$, on obtient un écart-type asymptotique inférieure, et qui diminue en fonction que la fréquence augmente.

	$n = 100$			$n = 200$			$n = 400$		
	g_1	g_2	g_3	g_1	g_2	g_3	g_1	g_2	g_3
s	0.00262	0.00495	0.00396	0.00099	0.00186	0.00139	0.00033	0.00059	0.00036
$n^{-\frac{3}{2}}\tau$	0.00240	0.00406	0.00186	0.00085	0.00143	0.00065	0.00030	0.00051	0.00023
$\frac{s}{n^{-\frac{3}{2}}\tau}$	1.09	1.22	2.13	1.17	1.29	2.11	1.10	1.17	1.57

TAB. 5.13 – Table des comparaisons de la statistique s et de l'écart-type asymptotique $n^{-\frac{3}{2}}\tau$ pour des échantillons de données de taille $n = 100, 200$. Les signaux considérés g_1 , g_2 et g_3 ont une fréquence $\omega_0 = 0.1$ et une amplitude $R = 2$. L'échantillonnage est un processus de Poisson $\mathcal{P}(1)$. Les erreurs sont modélisées par un processus CAR(1) avec $\phi = 0.5$.

5.4 Simulation de données avec le signal g_5 , avec erreurs gaussiennes indépendantes

Nous vérifions cette formule par des simulations avec un échantillonnage de Poisson $\mathcal{P}(1)$, avec un rapport signal bruit SNR = 2. Cette formule montre une indépendance de la fréquence du signal considéré. Pour vérifier cela, nous avons simulé le signal g_5 avec des fréquences $\omega_0 = 0.1, 0.2, 0.4, 0.8$, échantillonné avec un processus de Poisson $\mathcal{P}(1)$, sans ajout d'erreurs. La figure 5.6, en échelle logarithmique pour les deux axes, dont l'axe des ordonnées représente la valeur de la statistique s^2 en fonction du nombre d'observations $n = 100, 200, 300, 400$ représentés sur l'axe des abscisses, montre l'indépendance de la variance en fonction de la fréquence considérée.

Les valeurs de la statistique s ainsi que la variance théorique $\sqrt{\frac{1}{n^3}\tau_{g_5}} = \sqrt{\frac{1}{n^4}\text{Var}_{g_5}}$, pour le signal g_5 , de fréquence $\omega_0 = 0.1$, sont présentées dans le tableau 5.15.

		$n = 100$			$n = 200$			$n = 400$		
		g_1	g_2	g_3	g_1	g_2	g_3	g_1	g_2	g_3
$\omega_0 = 0.1$	s	0.00210	0.00404	0.00345	0.00076	0.00149	0.00109	0.00027	0.00048	0.00030
	$n^{-\frac{3}{2}}\tau$	0.00135	0.00225	0.00104	0.00047	0.00079	0.00036	0.00016	0.00028	0.00013
	$\frac{s}{n^{-\frac{3}{2}}\tau}$	1.55	1.79	3.31	1.60	1.87	2.96	1.61	1.71	2.32
$\omega_0 = 0.8$	s	0.00181	0.00337	0.00449	0.00063	0.00111	0.00116	0.00019	0.00037	0.00035
	$n^{-\frac{3}{2}}\tau$	0.00101	0.00222	0.00216	0.00036	0.00078	0.00076	0.00092	0.00027	0.00027
	$\frac{s}{n^{-\frac{3}{2}}\tau}$	1.78	1.51	2.08	1.76	1.41	1.52	1.54	1.33	1.30
$\omega_0 = 1.5$	s	0.00149	0.00291	0.00291	0.00055	0.00107	0.00102	0.00019	0.00035	0.00036
	$n^{-\frac{3}{2}}\tau$	0.00070	0.00206	0.00211	0.00024	0.00073	0.00074	0.00008	0.00025	0.00026
	$\frac{s}{n^{-\frac{3}{2}}\tau}$	2.12	1.41	1.37	2.23	1.47	1.37	2.22	1.37	1.37

TAB. 5.14 – Table des comparaisons de la statistique s et de l'écart-type asymptotique $n^{-\frac{3}{2}}\tau$ pour des échantillons de données de taille $n = 100, 200, 400$. Les signaux considérés g_1, g_2 et g_3 ont une fréquence $\omega_0 = 0.1, 0.8, 1.5$ et une amplitude $R = 2$. L'échantillonnage est un processus de Poisson $\mathcal{P}(1)$. Les erreurs sont modélisées par un processus CAR(1) avec $\phi = 0.9$.

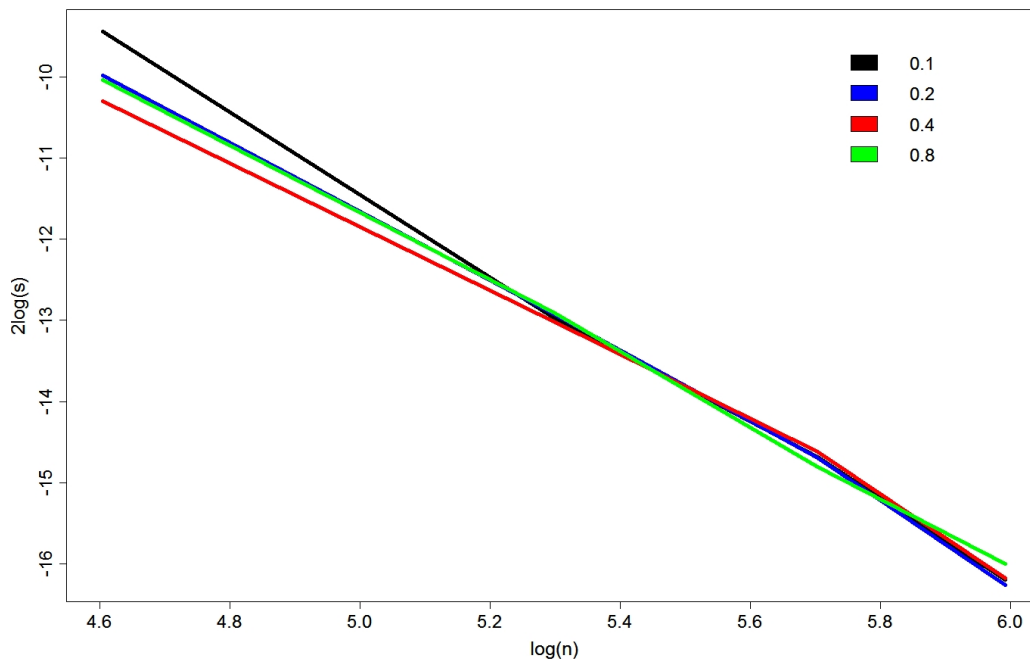


FIG. 5.6 – Echelle des axes log-log. Vérification de l'indépendance de la variance pour des fréquences différentes du signal g_5 . La pente obtenue est d'environ 4, ce qui rend crédible la formule de la variance (4.136).

	$n = 100$	$n = 200$	$n = 300$	$n = 400$
s	0.00062	0.00010	0.000045	0.000021
$\frac{1}{n^2} \sqrt{\text{Var}_{g_5}}$	0.00027	0.000068	0.000030	0.000017
$\frac{s}{\frac{1}{n^2} \sqrt{\text{Var}_{g_5}}}$	2.31	1.590241	1.516572	1.268102

TAB. 5.15 – Table des comparaisons de la statistique s et de la variance théorique pour des échantillon de données de taille $n = 100, 200, 300, 400$ pour le signal g_5 de fréquence $\omega_0 = 0.1$. L'échantillonnage est un processus de Poisson $\mathcal{P}(1)$.

Comme nous l'avons vu, la variance asymptotique de $\widehat{\omega}_n^{\text{NP}}$ est nettement meilleure que celle de $\widehat{\omega}_n^{\text{AH}}$, pour le g_5 . Même pour $n = 100$ et 200 observations, avec un processus de Poisson $\mathcal{P}(1)$ pour les temps d'observations, la méthode NP est supérieure à AH. Le tableau 5.16 ne montre pas de différence entre les valeurs de s_{AH} et s_{NP} . Cependant, les valeurs p_{NP} sont plus élevées que celles de p_{AH} . Ainsi l'estimateur $\widehat{\omega}_{\text{NP}}$ obtient une meilleure performance que $\widehat{\omega}_{\text{AH}}$.

	$n = 100$	$n = 200$
s_{AH}	0.0140	0.0070
p_{AH}	0.049	0.075
s_{NP}	0.0116	0.0067
p_{NP}	0.221	0.466

TAB. 5.16 – Table des statistiques s_{NP} des échantillons de données de taille $n = 100, 200$ pour le signal g_5 de fréquence $\omega_0 = 0.1$. Les valeurs de p_{NP} et p_{AH} sont les proportions d'estimations utilisées pour le calcul de s , et donc se trouvant dans un voisinage proche de la la fréquence ω_0 . L'échantillonnage est un processus de Poisson $\mathcal{P}(1)$, avec un rapport signal bruit de 2. Ici la fonction de régression est estimée par repliement des observations.

5.5 Comparaison de la variance asymptotique avec la variance estimée

Dans cette dernière section le tableau 5.17 présente des comparaisons entre la variance asymptotique et les variances estimées s_{NP} pour les signaux g_1, g_2, g_3 et g_4 sur un échantillonnage suivant un processus de Poisson $\mathcal{P}(1)$. Contrairement aux simulations effectuées plus haut, nous considérons comme connues la forme du signal, ainsi que son amplitude. Cela permettra d'avoir une meilleure estimation de l'écart-type estimé s_{NP} . Ainsi la variance asymptotique (4.13) est simplement divisée par 4. En effet, si le signal est connu, il ne reste plus que la fréquence comme paramètre à estimer. La variance asymptotique devient donc

$$\tau^2 = 2\pi\sigma^2 3\omega_0\eta^2 \left\{ \int_0^{P_0} g'(u)^2 du \right\}^{-1}.$$

La variance asymptotique (4.132) est aussi divisée par 4. Ici le symbole * signifie que la valeur est inférieure à 10^{-6} . Le tableau 5.17 montre une cohérence de la variance asymptotique avec la variance estimée, avec un plus grand écart pour le signal g_3 avec $n = 100$. On remarque que avec $n = 200$ cet écart diminue. Dans le tableau 5.18, on peut remarquer que les valeurs de s_{NP} sont simplement les valeurs de s_{NP} du tableau 5.17 divisées par environ 24 et ce pour les signaux g_1 et g_2 . En effet, le modèle d'échantillonnage $\mathcal{P}(1)$ a un $\eta = 1$ et pour le modèle d'échantillonnage EchNU₁, $\eta = \delta \simeq 24$. On vérifie donc bien les résultats que nous avons obtenus lors de l'analyse des résultats lors d'un échantillonnage EchNU₁ de type B et C dans la section 5.2.2. Par contre, en comparant les valeurs de

s pour le signal g_4 dans les tableaux où $n = 100$, on obtient des rapports qui peuvent être éloignés de δ , ce qui veut dire que les discontinuités de cette fonction ne sont pas si bien couvertes par ce modèle d'échantillonnage. Ces simulations ont relevé que la variance asymptotique n'est pas péjorée, par rapport à la variance asymptotique du cas d'échantillonnage vérifiant la condition de la couverture de la période, et ce pour toute valeur de $\frac{P_0}{2} > \delta > 0$, lorsque certaines parties du signal g_1 ou g_2 ne sont pas observées et que l'échantillonnage est un modèle EchNU₁. Borner δ par $\frac{P_0}{2}$ permet d'inclure aussi le cas $\beta = 0$, comme indiqué à la fin de la section 4.3.

		Ech ₂ , $n = 100$							
		SNR = 1				SNR = 2			
		g_1	g_2	g_3	g_4	g_1	g_2	g_3	g_4
$\omega_0 = 0.1$	p_{NP}	0.961	0.813	0.84	0.921	0.957	0.944	0.947	0.88
	s_{NP}	0.0026	0.0049	0.0042	0.0013	0.0013	0.0024	0.0018	0.0004
	τ_{NP}	0.0024	0.0040	0.0018	0.0004	0.0012	0.0020	0.0009	0.0002
$\omega_0 = 0.8$	p_{NP}	0.96	0.831	0.743	0.997	0.947	0.933	0.929	0.908
	s_{NP}	0.0026	0.018	0.0233	0.0061	0.0013	0.0024	0.0025	0.0004
	τ_{NP}	0.0024	0.0040	0.0041	0.0004	0.0012	0.0020	0.0020	0.0002
$\omega_0 = 1.5$	p_{NP}	0.957	0.86	0.742	0.99	0.941	0.918	0.958	0.883
	s_{NP}	0.0027	0.0483	0.0521	0.0152	0.0013	0.0025	0.0026	0.0004
	τ_{NP}	0.0024	0.0043	0.0043	0.0004	0.0012	0.0022	0.0022	0.0002

		Ech ₂ , $n = 200$							
		SNR = 1				SNR = 2			
		g_1	g_2	g_3	g_4	g_1	g_2	g_3	g_4
$\omega_0 = 0.1$	p_{NP}	0.956	0.933	0.931	0.96	0.957	0.945	0.939	0.999
	s_{NP}	0.0009	0.0017	0.0015	0.0002	0.0004	0.0008	0.0006	*
	τ_{NP}	0.0008	0.0014	0.0007	0.0001	0.0004	0.0007	0.0003	0.00005
$\omega_0 = 0.8$	p_{NP}	0.921	0.937	0.938	0.965	0.758	0.944	0.923	0.998
	s_{NP}	9e-4	0.0017	0.0030	0.0002	0.0005	0.0008	0.0009	*
	τ_{NP}	0.0008	0.0014	0.0014	0.0001	0.0004	0.0007	0.0007	0.00005
$\omega_0 = 1.5$	p_{NP}	0.914	0.98	0.94	0.952	0.755	0.932	0.933	0.998
	s_{NP}	0.0009	0.0209	0.0314	0.0002	0.0005	0.0009	0.0009	*
	τ_{NP}	0.0008	0.0015	0.0015	0.0001	0.0004	0.0007	0.0007	0.00005

TAB. 5.17 – Table des valeurs des proportions p_{NP} d'estimations correctes selon le critère CC₁, des valeurs de s_{NP} , ainsi que de l'écart-type asymptotique τ_{NP} , pour un échantillon de données de taille $n = 100$ et 200 provenant d'un processus de Poisson $\mathcal{P}(1)$. Les fréquences sont $\omega_0 = 0.1, 0.8$ et 1.5 .

Le tableau 5.20 donne les divers cas de figures couverts par les simulations.

5.6 Conclusion

Dans ce chapitre, nous avons comparé les performances des méthodes D, AH et NP sur des simulations de quatre signaux périodiques, avec trois différentes amplitudes et fréquences, échantillonnés avec des processus de temps vérifiant la définition 2 de couverture de la période, ainsi que des processus de temps ne la vérifiant pas. Nous avons considéré 100 et 200 observations. Globalement, le tableau 5.3 montre que pour les processus de temps vérifiant la définition 2, la méthode avec la meilleure efficacité est la méthode AH et cela pour les quatre signaux, lorsque le rapport signal bruit vaut 1, 2 ou 3. Dans l'ensemble, la méthode AH est non seulement plus efficace que la méthode NP, mais aussi beaucoup plus rapide. Nous avons aussi comparé les valeurs des écarts-types s_{NP} et s_{AH} avec les valeurs des écarts-types asymptotiques τ_{NP} et τ_{AH} respectivement. Pour la plupart des cas, il a été difficile d'obtenir une bonne comparaison entre s_{NP} et τ_{NP} . Nous avons rencontré trois cas de figure. Le premier étant qu'on obtenait des valeurs assez proches entre s_{NP} et τ_{NP} . Par exemple, pour le signal g_1 dans les tableaux 5.3 et 5.4 considérant des échantillonnages de processus uniforme et de Poisson. Le deuxième cas de figure étant que la valeur de s_{NP} n'est pas calculable (symbole *), exemple dans

EchNU₁, type B, n = 100								
SNR = 1					SNR = 2			
	g_1	g_2	g_3	g_4	g_1	g_2	g_3	g_4
p_{NP}	0.96	0.799	0.78	0.979	0.985	0.918	0.933	1
s_{NP}	0.0001	0.0009	0.001	*	0.0001	0.0001	0.0001	*
τ_{NP}	0.000104	0.000169	0.000058	-	0.000052	0.000084	0.000029	-

EchNU₁, type B, n = 200								
SNR = 1					SNR = 2			
	g_1	g_2	g_3	g_4	g_1	g_2	g_3	g_4
p_{NP}	0.963	0.975	0.971	1	0.811	0.967	0.928	1
s_{NP}	0.000040	0.000337	0.000235	*	0.000021	0.000037	0.000016	*
τ_{NP}	0.000036	0.000059	0.0000205	-	0.000018	0.000030	0.000010	-

TAB. 5.18 – Table des valeurs des proportions p d'estimations correctes selon le critère CC_2 , avec un intervalle de confiance $I_{n,i}$, basé sur l'écart-type τ , pour $\omega_0 = 0.1$. L'échantillon de données de taille $n = 100$, pour le premier tableau, et $n = 200$, pour le second tableau. Le modèle d'échantillonnage est le modèle **EchNU₁** de type B.

EchNU₁, type C, n = 100								
SNR = 1					SNR = 2			
	g_1	g_2	g_3	g_4	g_1	g_2	g_3	g_4
p_{NP}	0.919	0.754	0.75	0.966	0.976	0.904	0.907	1
s_{NP}	0.000108	0.000779	0.001318	0.0006	0.000055	0.000096	0.000082	0.000016
τ_{NP}	0.000103	0.000167	0.0000440	-	0.000052	0.000084	0.000022	-

EchNU₁, type C, n = 200								
$\frac{A}{\sigma} = 1$					$\frac{A}{\sigma} = 2$			
	g_1	g_2	g_3	g_4	g_1	g_2	g_3	g_4
p_{NP}	0.947	0.947	0.918	1	0.815	0.965	0.97	1
s_{NP}	0.000040	0.000286	0.000204	*	0.000021	0.000037	0.000025	*
τ_{NP}	0.000036	0.000059	0.0000155	-	0.000018	0.0000295	0.0000078	-

TAB. 5.19 – Table des valeurs des proportions p d'estimations correctes selon le critère CC_2 , avec un intervalle de confiance $I_{n,i}$, basé sur l'écart-type τ , pour $\omega_0 = 0.1$. L'échantillon de données de taille $n = 100$, pour le premier tableau, et $n = 200$, pour le second tableau. Le modèle d'échantillonnage est le modèle **EchNU₁** de type C.

Vérifiant la définition 2	Nom de l'échantillonnage	Erreurs
Oui	Ech ₁	iid
Oui	Ech ₂	iid
Oui	Ech ₃	iid
Oui	Ech ₄	iid
Oui	Ech ₂	Corrélées, CAR(1)
Non	EchNU ₁	iid
Non	EchNU ₂	iid

TAB. 5.20 – Cas couverts par les simulations

le tableau 5.5, avec $n = 200$, considérant des échantillonnages de mixture d'exponentielles, avec le signal g_1 et g_4 . Cela est dû à la grille des fréquences considérées utilisée dans les méthodes. Une grille plus fine aurait permis une meilleure comparaison, car une meilleure estimation de la variabilité des estimateurs $\hat{\omega}_n^D$, $\hat{\omega}_n^{AH}$ et $\hat{\omega}_n^{NP}$. Mais les temps de calculs n'auraient plus été raisonnables, surtout pour la méthode NP. Le dernier cas de figure est lorsque la valeur de s_{NP} est supérieure à τ_{NP} . Exemple dans le tableau 5.4, avec $n = 100$, pour le signal g_2 et g_3 . Le nombre de données n'étant pas forcément assez grand pour toujours obtenir une bonne estimation de la fonction de régression, ce qui donne une moins bonne estimation de la fréquence. C'est pourquoi, dans la dernière section 5.5 du chapitre, nous avons effectué quelques simulations avec une grille de fréquence plus fine et avec fonction de régression connue. Cela nous a permis d'obtenir une meilleure comparaison de s_{NP} et τ_{NP} et cela pour valider nos formules asymptotiques. Nous avons aussi comparé la performance de la méthode AH, la meilleure des méthodes, en utilisant l'intervalle de confiance asymptotique pour déterminer la proportion d'estimations correctes. Les résultats ont montré que suivant l'échantillonnage considéré, l'estimateur est plus au moins précis. On obtient une très grande précision lors du processus de temps avec mixture d'exponentielles dès que le rapport signal bruit vaut 2 ou 3. En effet, la proportion d'estimations couvertes par l'intervalle de confiance est proche de 1. La performance est plus élevée car l'espérance de la différence entre deux temps consécutifs η vaut 4.06. Comme nous l'avons vu, la variance asymptotique est divisée par η^2 . De plus, la grille des fréquences à tester par la méthode influence positivement ces résultats. Avec une grille plus fine, nous aurions toujours des très bons résultats, mais pas forcément des proportions égales à 1. Cette partie a aussi permis de vérifier la formule de la variance asymptotique de l'estimateur $\hat{\omega}_n^{AH}$ du théorème 9. Dans la deuxième section 5.2.2 de ce chapitre, nous avons comparé les performances des méthodes D, AH et NP sur des simulations de quatre signaux périodiques, avec trois différentes amplitudes et une fréquence. L'échantillonnage étant lié à la période et ne vérifiant pas la définition 2. La méthode AH reste la plus performante lorsque le modèle d'échantillonnage est EchNu₁ de type B ou C. En fait le type C se rapproche du cas régulier. Avec le modèle d'échantillonnage EchNU₂ de type A, les méthodes les plus performantes sont les méthodes D et NP. La méthode AH est nettement moins performante. Avec le modèle d'échantillonnage EchNU₂ de type B et C, la performance des trois méthodes diminuent. Le fait de n'observer qu'un quart ou un huitième du signal pose de réels problèmes à ces trois méthodes. Pour les mêmes raisons évoquées plus haut pour le cas de l'échantillonnage vérifiant la définition 2, dans la section 5.5 nous avons aussi effectué quelques simulations avec une grille de fréquence plus fine et avec fonction de régression connue. Cela pour les modèles de temps EchNu₁ de type B et C, pour pouvoir comparer s_{NP} et τ_{NP} , et valider la formule de la variance asymptotique de l'estimateur $\hat{\omega}_n^{NP}$ du théorème 12. La section 5.4 a été consacrée à la vérification de la variance asymptotique de l'estimateur $\hat{\omega}_n^{NP}$ pour le signal g_5 , obtenue à la section 4.4.1. Nous avons pu vérifier l'indépendance de la formule avec la fréquence ainsi que la convergence d'ordre $O(n^{-4})$. La section 5.3 a permis de tester la formule de la variance asymptotique de la fréquence $\hat{\omega}_n^{NP}$ du théorème 11.

Chapitre 6

Conclusion

L'objectif de ce travail était de tester et d'adapter diverses méthodes de recherche de fréquences inconnues appliquées à des séries temporelles d'observations d'étoiles variables. De plus, ces méthodes se devaient d'être automatiques, sans intervention humaine. Nous avons tout d'abord considéré les méthodes classiques de la théorie des séries temporelles, comme diverses méthodes basées sur l'autocorrélation ainsi que le périodogramme en amplitude. Nous avons aussi considéré deux méthodes de régression, l'une avec une fonction de régression sinusoidale, l'autre avec une fonction de régression estimée à partir des observations.

Nous avons sélectionné les méthodes qui ont donné les meilleurs résultats lors de l'application de ces dernières sur les données de la partie A du catalogue des étoiles variables de la mission Hipparcos. Les méthodes sélectionnées ont été le périodogramme en amplitude et les deux méthodes de régressions, étant donné que les performances de la méthode basée sur l'autocorrélation sont vraiment trop mauvaises. La forme de la courbe lumière observée peut donner des difficultés aux méthodes basées sur les fonctions sinusoidales, comme par exemple les binaires à éclipses, contenant une composante sinusoidale de fréquence ω_0 faiblement présente.

Dans un deuxième temps, au chapitre 4 nous avons procédé à l'étude de la distribution asymptotique des estimateurs $\hat{\omega}_n^{\text{AH}}$ et $\hat{\omega}_n^{\text{NP}}$, pour le cas d'un signal g périodique quelconque de période P_0 , tel que $g|_{[0, P_0)} \in L^2([0, P_0))$, avec processus Gaussien stationnaire corrélé représentant les erreurs additives dans les observations. Nous avons démontré que la convergence de l'estimateur $\hat{\omega}_n^{\text{AH}}$ reste en $o_p(n^{-1})$, comme l'avait déjà démontré Walker [32] et [33] dans le cas d'observations régulières avec un signal sinusoidal. Nous avons aussi obtenu la distribution asymptotique de l'estimateur de $\hat{\omega}_n^{\text{AH}}$ pour le cas d'observations irrégulières vérifiant la condition de couverture de la période, avec signal g périodique de période P_0 , vérifiant les conditions (4.2), (4.3) et (4.4), et erreurs Gaussiennes corrélées. Il s'est avéré que la distribution est identique au cas d'un signal sinusoidal avec observations régulières [33]. Nous avons aussi obtenu la distribution asymptotique de l'estimateur de $\hat{\omega}_n^{\text{NP}}$ pour le cas d'observations irrégulières vérifiant la condition de couverture de la période, avec signal g périodique de période P_0 , vérifiant les conditions (4.2) et (4.3), et erreurs Gaussiennes corrélées.

Nous nous sommes aussi intéressés à la distribution asymptotique de l'estimateur de $\hat{\omega}_n^{\text{NP}}$ pour le cas d'observations irrégulières ne vérifiant pas la condition de couverture de la période, avec signal g périodique quelconque de période P_0 , vérifiant les conditions (4.2) et (4.3), et avec erreurs indépendantes. Finalement, nous nous sommes penchés sur la variance asymptotique de $\hat{\omega}_n^{\text{NP}}$ pour le cas d'un signal particulier g_5 comportant des discontinuités. Nous en avons dérivé un résultat qui montre que la variance asymptotique est d'ordre $O(n^{-4})$. Asymptotiquement, l'estimation de la fréquence de ce type de fonction est très bien estimée.

Finalement, au chapitre 5 nous avons effectué toute une série de simulations avec divers modèles d'échantillonnages. Une des ces catégories permet une bonne estimation de la forme du signal, c'est-à-dire que ces échantillonnages vérifient la condition de la couverture de la période. L'autre catégorie de modèles d'échantillonnages ne vérifie pas cette condition. Nous avons ainsi pu comparer les méthodes de Deeming, de l'analyse harmonique ainsi que la régression non paramétrique. Nous avons considéré quatre signaux périodiques avec trois fréquences et rapport signal/bruit différents. Le nombre d'observations a été fixé à 100 et 200. Ces séries de simulations ont montré que lorsqu'on considère des modèles d'échantillonnages vérifiant la couverture de la période, avec un échantillon d'observations de 100 ou 200 données, la méthode de l'analyse harmonique est plus performante que les deux autres. Mais nous avons aussi noté qu'un échantillonnage de type Neyman-Scott n'est pas favorable pour la méthode NP.

En ce qui concerne un échantillonnage ne vérifiant pas la couverture de la période, l'analyse harmonique et la méthode NP ont des performances similaires lorsque le modèle d'échantillonnage est EchNu₁ de type B ou C. En fait le type C se rapproche du cas régulier. Avec le modèle d'échantillonnage EchNU₂ de type A, la performance de la méthode diminue, mais les méthodes NP et D restent bonnes et supérieures à AH. Tandis pour les types B et C, les performances des trois méthodes sont plutôt basses. Sur le plan de la performance, en terme de temps de calculs, la méthode NP est plus coûteuse, de par le nombre d'opérations à effectuer. Repliement en phase des données pour chaque fréquence à tester, lissage des données et régression avec la fonction estimée. Ainsi, au vu des bons résultats obtenus par l'analyse harmonique, et de sa plus grande rapidité en terme de temps de calculs, on peut en conclure que cette méthode nettement plus avantageuse.

Nous avons effectué diverses simulations pour pouvoir comparer la variance estimée avec la variance asymptotique, dans le cas d'erreurs indépendantes et corrélées, ainsi que pour les signaux g_4 et g_5 , donnant une vitesse de convergence de la variance estimée vers la variance asymptotique plus rapide que pour les signaux continus. Pour obtenir une bonne estimation de la variance estimée, nous avons considéré les fonctions de régressions comme étant connues, avec un modèle d'échantillonnage vérifiant la condition de couverture de la période, et un autre modèle ne vérifiant pas cette condition. Ces simulations nous ont permis de mettre en lumière le fait que la variance asymptotique (4.132), pour le cas de l'échantillonnage EchNU₁ ne vérifiant pas la condition de couverture de la période, est en fait égale à la variance asymptotique (4.13) avec un $\eta = \delta$, pour le signal g_1 , et est très proche pour le signal g_2 . Ceci montre que, asymptotiquement et pour les catégories de fonctions dont la composante sinusoidale de fréquence ω_0 est prédominante, la variance asymptotique de l'estimateur de la fréquence est proche de la variance asymptotique lorsque les observations recouvrent l'intervalle de la période $[0, P_0)$. De nouvelles recherches permettraient d'explorer d'autres types d'échantillonnages ne vérifiant pas la condition de couverture de la période, ainsi que d'autres types de signaux, pour en dériver la variance asymptotique de l'estimateur de la fréquence.

Chapitre 7

Appendix

Théorème 13 (Distribution asymptotique de $\widehat{\omega}_n^{\text{NP}}$ avec la fonction périodique g connue)

Soit le modèle (4.14), avec erreur $\varepsilon(t)$, un processus stochastique Gaussien stationnaire et ergodique corrélié, avec fonction d'autocovariance $\gamma(s)$ vérifiant (4.5). Supposons que la densité spectrale $f_{sp}(\omega)$ de l'erreur existe. La fonction $g(t)$ est périodique de période P_0 vérifiant les conditions (4.2) et (4.3). Soit $\widehat{\omega}_n^{\text{NP}}$, l'estimateur de la fréquence ω_0 , défini à la section 2.1. L'échantillonnage est irrégulier, mais vérifiant la définition 2 de la couverture de la période. De plus, $\eta = \lim_{i \rightarrow \infty} \mathbb{E}[T_i - T_{i-1}]$ existe. Alors

$$n^{\frac{3}{2}} (\widehat{\omega}_n^{\text{NP}} - \omega_0) \sim \mathcal{N}(0, \tau^2) \quad (7.1)$$

où

$$\tau^2 = 3\omega_0 \left(\int_0^{P_0} (g'(v)dv)^2 \right)^{-2} \int_{-\infty}^{+\infty} \left| \frac{\partial}{\partial \omega} \mathcal{TF} \{g'\}(\omega) \right|^2 f(\omega) d\omega \quad (7.2)$$

Preuve. Nous considérons que la fonction g est connue. Ainsi il suffit de reprendre la preuve du théorème 11, après l'équation (4.116), ce qui va conduire à obtenir un facteur $\frac{1}{3}$ au lieu de $\frac{1}{12}$ dans le calcul de la variance, ce qui donne

$$\text{Var} [\widehat{\omega}_n^{\text{NP}} - \omega_0] \simeq \frac{1}{3} \omega_0 \frac{1}{n^3} \left(\frac{1}{3} \int_0^{P_0} g'(u)^2 du \right)^{-2} \int_{-\infty}^{+\infty} |\mathcal{TF} \{g'\}(\omega)|^2 f(\omega) d\omega$$

■

Bibliographie

- [1] P. Bloomfield. Spectral analysis with randomly missing observations. *Journal of the Royal Statistical Society, Series B*, 32 :369–380, 1970.
- [2] P. Bloomfield. *Fourier Analysis of Time Series : An Introduction*. Wiley Series in Probability and Mathematical Statistics, 1976.
- [3] P. J. Brockwell and R. A. Davis. *Time Series : Theory and Methods*. Springer Series in Statistics, 1987.
- [4] D. R. Cox and H. D. Miller. *The Theory of Stochastic Processes*. Science Paperback, 1977.
- [5] T. J. Deeming. Fourier analysis with unequally-spaced data. *Astrophysics and Space Science*, 36 :137–158, 1975.
- [6] R. A. Edelson and J. H. Krolik. The discrete correlation function : a new method for analysing unevenly sampled variability data. *The Astrophysical Journal*, 333 :646–659, 1988.
- [7] L. Eyer and P. Bartholdi. Variable stars : Which Nyquist frequency? *Astronomy and Astrophysics Supplement*, 135, 1999.
- [8] S. Ferraz-Mello. Estimation of periods from unequally spaced observations. *Astronomical Journal*, 86, 1981.
- [9] R. A. Fisher. Tests of significance in harmonic analysis. *Proceedings of the Royal Society of London. Series A*, 125 :54–59, 1929.
- [10] G. Foster. The Cleanest Fourier Spectrum. *Astronomical Journal*, 109, 1994.
- [11] C. M. Gaskell and B. M. Peterson. the accuracy of cross-correlation estimates of quasar emission-line region sizes. *Astrophysical Journal Supplement Series*, 65 :1–11, 1987.
- [12] C. M. Gaskell and L. S. Sparke. Line variations in quasars and Seyfert galaxies. *Astrophysical Journal*, 305 :175–186, 1986.
- [13] P. Hall, J. Reimann, and J. Rice. Nonparametric estimation of a periodic function. *Biometrika*, 87, 2000.
- [14] R. H. Jones. Spectral analysis with regularly missed observations. *The Annals of Mathematical Statistics*, 33 :455–461, 1962.
- [15] A. M. Kovacs. Frequency shift in Fourier analysis. *Astrophysics and Space Science*, 78 :175–188, 1981.
- [16] G. Kovacs, S. Zucker, and T. Mazeh. A box-fitting algorithm in the search for periodic transits. *Astronomy and Astrophysics*, 391 :369–377, 2002.
- [17] N. R. Lomb. Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science*, 39 :447–462, 1976.
- [18] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1998.
- [19] D. W. Marquardt and S. K. Acuff. Direct Quadratic Spectrum Estimation with Irregularly Spaced Data. *Lecture Notes in Statistics, Springer-Verlag*, 1983.
- [20] R. J. Martin. Autoregression and irregular sampling : Filtering. *Signal Processing*, 69 :229–248, 1998.

- [21] P. Martinez and C. Koen. Period searching by least squares in sparsely sampled light curves with non-sinusoidal oscillations. *Royal Astronomical Society*, 267 :1039–1044, 1994.
- [22] J. A. McDonald. Periodic smoothing of time series. *SIAM J.Sci. Statist Comp.*, 7 :665–688, 1986.
- [23] F. Mignard. About the Nyquist frequency. Technical report, Observatoire de la Côte d’Azur, 2005.
- [24] S. Morgenthaler. *Introduction à la statistique*. Presses polytechniques et universitaires romandes, 1997.
- [25] D. H. Roberts, J. Lehář, and J. W. Dreher. Time Series With Clean. i. Derivation of a Spectrum. *The Astronomical Journal*, 93, 1987.
- [26] J. D. Scargle. Studies in astronomical time series analysis ii. Statistical aspects of spectral analysis of unevenly spaced data. *Astrophysical Journal*, 263 :835–853, 1982.
- [27] A. Schuster. On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena. *Terrestrial Magnetism and Atmospheric Electricity*, 3 :13–41, 1898.
- [28] A. Schwarzenberg-Czerny. On the advantage of using analysis of variance for period search. *Royal Astronomical Society*, 241 :153–165, 1989.
- [29] H. S. Shapiro and R. A. Silverman. Alias-free sampling of random noise. *SIAM*, 8, 1960.
- [30] P. Vaniček. Approximate spectral analysis by least squares fit. *Astrophysics and Space Science*, 4 :387–391, 1969.
- [31] P. Vaniček. Further development and properties of the spectral analysis by least squares. *Astrophysics and Space Science*, 12 :10–33, 1970.
- [32] A. M. Walker. On the estimation of a harmonic component in a time series with stationary independent residuals. *Biometrika*, 58 :21–36, 1971.
- [33] A. M. Walker. On the estimation of a harmonic component in a time series with stationary dependent residuals. *Advances in Applied Probability*, 5 :217–241, 1973.
- [34] R. J. White and B. M. Peterson. Comments on Cross-Correlation Methodology in Variability Studies of Active Galactic Nuclei. *Publications of the Astronomical Society of the Pacific*, 106 :879–889, 1994.
- [35] M. Zechmeister and M. Kürster. The generalised Lomb-Scargle periodogram, a new formalism for the floating-mean and Keplerian periodograms. *Astronomy and Astrophysics Supplement*, 496 :577–584, 2009.

Jean-Marc Nicoletti
11, Rue de la Mèbre
1033 Cheseaux-sur-Lausanne
076 588 52 22

22 mai 1979
Nationalité suisse
jm.nicoletti@hotmail.com

DOCTEUR EN STATISTIQUES

Expérience dans le traitement et l'analyse de données ainsi que la programmation

DOMAINE DE COMPÉTENCES :

- Analyses statistiques, recherche de périodicité dans un jeu de données, développement et implémentation de méthodes robustes, recherche du comportement asymptotique d'un estimateur statistique.
- Participation au projet GAIA de l'Agence Spatiale Européenne, en collaboration avec l'observatoire de Genève : Recherche de périodes des étoiles variables (depuis 2007).
- Mathématiques appliquées : modélisation d'un réseau d'eau potable, implémentation d'algorithmes en Java, "débruitage" d'un signal par une décomposition en ondelette.

PARCOURS PROFESSIONNEL :

06.2007 – ce jour

EPFL – Doctorant et Assistant en statistiques

- Spécialités : statistiques, probabilités, mathématiques.
- Enseignements : Introduction aux probabilités et aux statistiques, analyse, mathématiques générales, géométrie.
- Encadrement : 200 étudiants ; divers travaux de semestre (méthodes de classification, recherche de périodes).

11.2006 – 04.2007

SOURCINGPARTS, Coppet - Analyste développeur

- Développement d'outils d'administration internes en langage Java et Flex. Maintenance des méthodes (mise à jour, traitement des "bugs"). Participation au développement de la nouvelle plateforme en Flex.

05.2006

FISCHER CONNECTORS, Apples – Aide de laboratoire

- Essais en laboratoire de qualification d'un nouveau connecteur. Elaboration d'un petit guide de l'utilisateur du programme statistique StatGraphics. Rédaction de rapports de laboratoire.

FORMATION :

2012

Doctorat en statistiques – EPFL.

2006

Master en ingénierie mathématique – EPFL.

1998

Baccalauréat et maturité en Sciences économiques.

LANGUES :

Français, Italien :

langues maternelles.

Anglais :

courant.

Allemand :

compréhension orale et écrite.

CONNASSANCES INFORMATIQUES :

Logiciels spécialisés : MatLab ; R ; Environnement Eclipse ; programmation en Java.

ACTIVITES EXTRAPROFESSIONNELLES :

Cyclisme, course à pied, kick-boxing. Intérêts pour la météorologie, l'astronomie, les sciences, les voyages, la cuisine et la musique.

PUBLICATION:

Antoine Musitelli, Jean-Marc Nicoletti. Competitive ratio of List Scheduling on uniform machines and randomized heuristics. *Journal of Scheduling*, volume 14, 2011, page 89.