



HAL
open science

Statistical methods using hydrodynamic simulations of stellar atmospheres for detecting exoplanets in radial velocity data

Sophia Sulis

► **To cite this version:**

Sophia Sulis. Statistical methods using hydrodynamic simulations of stellar atmospheres for detecting exoplanets in radial velocity data. Other. COMUE Université Côte d'Azur (2015 - 2019), 2017. English. NNT: 2017AZUR4073 . tel-01687077

HAL Id: tel-01687077

<https://theses.hal.science/tel-01687077>

Submitted on 18 Jan 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ CÔTE D'AZUR
École Doctorale en Sciences Fondamentales et Appliquées

Thèse de doctorat

Présentée en vue de l'obtention du
grade de docteur en
Sciences de l'Univers

au

Laboratoire J.-L. Lagrange - UMR 7293
Observatoire de la Côte d'Azur

par

Sophia SULIS

Méthodes statistiques utilisant des simulations hydrodynamiques
d'atmosphères stellaires pour détecter des exoplanètes en vitesse radiale

-

Statistical methods using hydrodynamic simulations
of stellar atmospheres for detecting exoplanets in radial velocity data

Dirigée par **David Mary**, professeur à l' Université de Nice Sophia Antipolis,
et codirigée par **Lionel Bigot**, chargé de recherche au CNRS au Laboratoire Lagrange.

Soutenue le 2 octobre 2017 devant la commission d'examen:

Rapporteurs:	Nadège Meunier	-	Astronome	-	IPAG, Grenoble
	Olivier Michel	-	Professeur	-	INPG, Grenoble
Examineurs:	François Bouchy	-	Professeur associé	-	Université de Genève, Suisse
	Rémy Boyer	-	Maître de conférences	-	L2S, Centrale Supélec
	Magali Deleuil	-	Professeur	-	Université de Marseille
	Maria Süveges	-	Chercheuse associée	-	Max-Planck Institute, Allemagne

Résumé:

Quand le bruit affectant une série temporelle est coloré avec des propriétés statistiques inconnues, la difficulté pour la détection de signaux périodiques est de contrôler le degré de confiance avec lequel les tests de détection rejettent l'hypothèse nulle en faveur de l'hypothèse alternative. Cette thèse développe la possibilité d'utiliser des séries temporelles simulées du bruit pour améliorer ce contrôle.

Plus particulièrement, dans le cas d'un échantillonnage régulier des observations, nous avons analysé les performances de différents tests de détection appliqués à un périodogramme standardisé par les séries temporelles simulées du bruit coloré. Notre étude a notamment été portée à considérer le cas de détections dites *parcimonieuses* dans le domaine de Fourier ainsi qu'à étudier l'influence du nombre, nécessairement fini, de simulations disponibles du bruit sur les performances de tests de détection appliqués à ce périodogramme standardisé. Nous avons dérivé des expressions analytiques pour les probabilités de fausses alarmes et de détections de différents tests et montré que la standardisation proposée entraîne, dans la majorité des cas, des tests de détection puissants dont le taux de fausses alarmes est constant (dits *CFAR*). Bien que ces résultats analytiques ont été dérivés dans le cadre d'un régime asymptotique, les résultats numériques montrent que la théorie développée décrit précisément le comportement des tests de détection pour un nombre modéré de points d'observation.

Dans le cas d'un échantillonnage irrégulier des observations, le taux de fausses alarmes ne peut pas être dérivé par des expressions analytiques simples même dans le cas d'un bruit blanc Gaussien. Cependant, nous montrons qu'il est possible de combiner la standardisation proposée du périodogramme avec des techniques de bootstrap pour contrôler la probabilité de fausses alarmes de manière fiable dans le cas général d'un bruit coloré dont on est capable de générer des séries temporelles. Nous montrons aussi que la procédure peut être améliorée en utilisant la distribution des valeurs extrêmes généralisées.

Tout au long de cette étude, nous nous intéressons au problème particulier de la détection de planètes extrasolaires par la méthode des vitesses radiales (VR). Avec la précision instrumentale des spectrographes actuels, la principale barrière pour détecter des planètes de masse terrestre vient de l'activité de l'étoile hôte elle-même: tâches, plages, convection et oscillations stellaires. Cette activité entraîne des corrélations dans les VR observées qui peuvent masquer la signature de planètes extrasolaires ou au contraire les imiter et produire de fausses détections. Alors que des indicateurs chromosphériques externes peuvent (partiellement) aider à supprimer les effets (quasi-) périodiques d'origine magnétique (comme l'effet des tâches), cette thèse s'intéresse à la contribution des processus de convection. La convection à la surface des étoiles induit un "jitter" d'origine stellaire permanent dans les VR observées dont l'amplitude est de l'ordre de celles des signatures de petites planètes. En parallèle, des simulations hydrodynamiques (*ab initio*) de l'activité de convection stellaire ont été développées depuis maintenant plusieurs décennies et permettent de reproduire, de manière fiables, des séries temporelles de ce "bruit" convectif. Notre objectif est de prouver que de telles simulations peuvent améliorer, de manière statistique le contrôle des seuils de fausses alarmes associés à toute détection de planètes. Dans cet esprit, une comparaison de données solaires en VR observées et simulées est présentée à la fin de ce manuscrit afin de confirmer la fiabilité de telles simulations. Dans ce contexte, la standardisation proposée du périodogramme et les tests de détections étudiés ouvrent la porte à de nombreuses applications intéressantes pour le domaine de la détection de planètes par la méthode des VR.

Mots clés: Traitement du signal, Statistique de tests, Exoplanètes, Vitesses radiales, Bruit stellaire coloré, Détection parcimonieuse.

Abstract:

When the noise affecting time series is colored with unknown statistics, a difficulty for periodic signal detection is to control the true significance level at which the detection tests are conducted. This thesis investigates the possibility of using training datasets of the noise to improve this control.

Specifically, for the case of regularly sampled observations, we analyze the performances of various detectors applied to periodograms standardized using the noise training datasets. Emphasis is put on *sparse* detection in the Fourier domain and on the limitation posed by the necessarily finite size of the training sets available in practice. We study the resulting false alarm and detection rates and show that the proposed standardization leads, in some cases, to powerful constant false alarm rate tests. Although analytical results are derived in an asymptotic regime, numerical results show that theory accurately describes the tests' behavior for moderately large sample sizes.

In the case of irregularly sampled observations, while analytical expressions for the false alarm rate are out of reach, we show that it is possible to combine the proposed periodogram standardization and bootstrap techniques to consistently estimate the false alarm rate. We also show that the procedure can be improved by using generalized extreme value distributions.

Throughout the study, we focus on the particular problem of extrasolar planet detection in radial velocity (RV) data. With the instrumental precision of the recent spectrographs, the main barrier to detect Earth-mass planets comes from the host star activity: e.g. spots and plages, convection, and stellar oscillations. This activity affects the RV as a colored noise, which can mimic or hide the planet signatures. While external chromospheric indicators can (partially) deal to avoid the (quasi-) periodic noise source due to the magnetic activity (e.g. the rotationally modulated starspots), this thesis focuses on the contribution of convective process, which induced a permanent stellar “jitter” leading to correlated RV variations with magnitude at the level of the smallest planet signatures. In parallel, reliable (*ab initio*) hydrodynamic simulations of the stellar surface convection activity have recently been developed during the last decade and we aim to prove that they can statistically improve the control of the significance assigned to planet detections. In this spirit, a comparison of RV solar observations and simulations will be presented at the end of this thesis to confirm the reliability of these simulations. In this context, the proposed periodogram standardization and detection tests investigated in the theoretical part of the thesis open several interesting applications in the field of exoplanet detection in RV data.

Keywords: Signal processing, Test statistics, Exoplanet, Radial velocity, Stellar colored noise, Sparse detection.

Remerciements:

Depuis toute petite, mon regard s'est toujours porté vers les étoiles et j'ai toujours espéré pouvoir passer ma vie à les regarder, à les étudier, et à pouvoir, à mon tour, apporter une petite pierre à l'édifice de notre connaissance de l'univers. Cette thèse ne s'est pas écrite que par ma plume mais est le fruit de la contribution de plusieurs personnes qui m'ont accompagnée, conseillée, soutenue et que je tiens à remercier chaleureusement car, sans elles, ce travail n'aurait jamais pu voir le jour.

En premier lieu, je tiens à remercier mes deux directeurs de thèse: David Mary et Lionel Bigot. David, merci pour ta patience, pour tous tes conseils, pour ta rigueur, et pour tout ce que tu m'as appris. Au début de ma thèse, j'étais une grande novice dans le domaine du traitement du signal et tu as su m'enseigner et me guider dans ce domaine avec patience et passion.

Lionel, merci également pour tes conseils, tes encouragements et ton enthousiasme quotidien. Je ne compte plus le nombre d'heures que tu as passé à me faire répéter et à revoir mes présentations. Vous avez su tout les deux me guider, me dire les choses qui allaient et celles qui allaient moins avec justesse et je n'oublierais pas les précieux conseils que vous m'avez donnés. Grâce à vous j'ai pu réaliser cette thèse dans les meilleures conditions possibles, j'ai pu voyager, apprendre des choses passionnantes et aussi murir, merci. C'était sincèrement un grand plaisir de pouvoir travailler avec vous et j'espère que l'on pourra continuer à collaborer dans le futur.

Je tiens aussi à exprimer ma gratitude à chacun des membres de mon jury, qui ont apporté un regard extérieur à ma thèse à la fois en traitement du signal et en astrophysique. Tout d'abord Nadège Meunier et Olivier Michel qui ont eu le courage d'être rapporteurs de ma thèse, et de la lire durant le mois d'août, entre vacances et conférences. Ensuite, je tiens à remercier Magali Deleuil qui a accepté d'être présidente du jury ainsi que Maria Süveges, Rémy Boyer et François Bouchy qui ont accepté d'endosser le rôle d'examineurs. J'en profite par ailleurs pour remercier une fois encore François Bouchy pour m'avoir accueillie à l'Observatoire de Genève et m'avoir permis de rencontrer son équipe.

Ce travail s'est fait dans un environnement idéal, aux côtés des chercheurs de l'Observatoire de la Côte d'Azur et notamment de l'équipe de physique stellaire et de traitement du signal que je remercie vivement. Merci donc à tous mes collègues: Frédéric pour avoir assisté à toutes mes répétitions et m'avoir accepté en tant qu'assistante guide de l'Observatoire, Catherine pour ton soutien durant ces trois années (et pour tous ces covoiturages et tes délicieux gateaux!), Olga pour m'avoir permis de faire tous ces ateliers avec les enfants et m'avoir appris à transmettre l'astronomie aux plus petits, Fabrice pour ces parties de ping-pong et ces pauses cafés, Delphine, Rita et Roula pour m'avoir enrôler dans la magnifique aventure de l'organisation du Gretsï, Rémy pour ton humour extraordinaire - ne le perds jamais - , Stephane pour ta visite en VIP et tes recettes de pâtes à l'italienne, Sergey pour tes conseils en littérature - russe comme française, Rabah pour ta gentillesse inégalée, Andrea, Orlagh, David, Janine, Thierry, FiX pour vos encouragements et vos conseils, mais aussi Fatme, Farrokh, Eric, Bertrand, Ali, Christine, Sylvie, Rose, Mamadou, Nicolas, Aurélie, Anthony, André, Claude, François, Marco et Aurélien pour votre bonne humeur quotidienne et votre accueil chaleureux. Merci également à Philippe Stee, Vanessa Hill, Paolo Tanga et Céline Theys pour le suivi de ma thèse au sein du laboratoire.

Ma thèse a bénéficié du soutien financier de la région PACA et de Thalès Alenia Space, que je tiens aussi à remercier. Ainsi que notamment Olivier Pigouche, Marion Neveu Van Malle et Bruno Valat pour leur présence et soutien en tant que référents Thalès.

Merci à Alain Lecavelier des Etangs et Vincent Bourrier pour m'avoir introduit à la thématique passionnante des planètes extrasolaires, pour m'avoir accompagnée et soutenue depuis mon master jusqu'à mon premier postdoc ainsi que Jean-Yves Chauffray avec qui j'ai eu l'occasion de travailler sur l'atmosphère de Vénus il y a quelques années.

Merci à mes amis de toujours: Cyndie, mon Compagnon depuis 23 ans, tu as toujours été là pour moi, dans toutes épreuves, main dans la main nous avons avancé ensemble. Je n'aurais jamais eu le courage de faire tout cela sans toi. Tu es mon amie et le sera éternellement. Jennifer, malgré les kilomètres qui nous séparent, tu fais partie intégrante de ma vie, et les moments que nous partageons ensemble sont pour moi inestimables.

Merci à mes vieux amis Benoît, Pierre, Raph, Thomas, Théo, Willy, Terry, Médé, Guillaume et Simon pour votre amitié depuis toutes ces années. Merci à Jocelyne et Marie de m'avoir hébergé lors de ma venue sur Nice et à mes amis niçois Muriel et David pour m'avoir fait découvrir cette région magnifique. Merci à mes collègues astrophysiciens de Paris pour tous ces bons moments et tout ces partages d'expériences. Et merci à Meg, mon ancienne professeure d'anglais, qui a eu le courage de me relire et qui m'a offert mon premier livre d'astrophysique en anglais que je garde avec moi de déménagements en déménagements.

Je voudrais enfin remercier ma famille qui m'a toujours soutenue dans ma scolarité alors que ce n'était pas facile. Tout d'abord ma mère, qui a toujours été un pilier dans ma vie. Merci de m'avoir donné tout ce que tu m'as donné sans jamais rien attendre en retour. Merci à mon grand frère pour tous tes appels de soutien et pour avoir toujours été là pour moi. Merci à mon père qui m'a appris que les seules barrières que l'on avait dans la vie étaient celles que l'on se mettait à soi-même. Et puis finalement, je remercie mon Rémy qui, entre temps, est devenu mon mari. Tu m'as toujours soutenue même dans les moments difficiles, quand je suis partie au Canada, à Paris, à Nice, lors de ma rédaction de thèse: tu as toujours été là. Je ne compte pas le nombre de nuits où tu es resté réveillé pour moi, pour me faire du thé, du café, et m'aider à tenir. Ta bonne humeur, ta patience et ton soutien sans faille sont pour moi une source de motivation inépuisable.

Graz, Autriche, décembre 2017

“All we have to decide is what to do with the time that is given us”

*The Fellowship of the Ring,
J.R.R. Tolkien, 1954*

Contents

Notations and Definitions	i
General Overview	v
List of Figures	viii
List of Tables	viii
I Introduction	1
1 Exoplanet detection	3
1.1 Methods and discoveries	4
1.2 The radial-velocity technique	10
1.2.1 Planetary signatures	10
1.2.2 Method of measurement and instrumental errors	13
1.3 Stellar noises	16
1.3.1 Magnetic activity	17
1.3.2 Oscillations	20
1.3.3 Convection	21
1.3.4 Summary	25
1.4 Simulations of the convective stellar activity	25
1.5 Conclusions on the astrophysical context of the thesis	27
2 Detection of periodic signatures in noise: tools, models and tests	29
2.1 Hypothesis testing	30
2.1.1 Design of hypotheses	30
2.1.2 Statistical tests and decision rule	31
2.1.3 Frequentist approaches: tests based on likelihood functions	38
2.1.4 A brief comment on Bayesian approaches	40
2.2 Periodogram(s)	40
2.2.1 The classical periodogram	40
2.2.2 Periodogram of a single sinusoid	41
2.2.3 Parseval's identity	44
2.2.4 Periodogram resolution	45
2.2.5 Bias and variance of the periodogram	45
2.2.6 Asymptotic distribution in the case of a pure random process	45
2.2.7 Asymptotic distribution of P in the case of a periodic signal in noise	46
2.2.8 A note on some modified periodograms	51
2.3 Planetary signatures in the periodogram	52
2.3.1 Signatures in the time domain	52
2.3.2 Signatures in the Fourier domain	53
2.3.3 Discussion and signal model	57
2.4 Statistical tests applicable to the periodogram	58
2.4.1 Preliminary notations	58

2.4.2	Classical tests	59
2.4.3	Tests designed for multiple sinusoids	62
2.5	Adaptive tests that can be applied to the periodogram	66
2.5.1	Introduction of the problem	66
2.5.2	Example: Kolmogorov-Smirnov test	66
2.5.3	Higher Criticism	67
2.5.4	Berk-Jones test	70
2.6	Sinusoid detection in unknown noise statistics: related works	71
2.6.1	Non parametric approaches	72
2.6.2	Parametric methods	74
2.7	Conclusions	75
2.7.1	Main conclusions	75
2.7.2	This thesis	76

II Proposed detection method 79

3 Statistical tests exploiting simulations of the colored noise: theoretical analysis and numerical study 81

3.1	Statement, assumptions and objectives of the detection problem	82
3.1.1	Model under both hypotheses	82
3.1.2	Assumptions and objectives	83
3.2	Statistical distribution of the proposed standardized periodogram	84
3.2.1	Standardized periodogram $\tilde{\mathbf{P}} \overline{\mathbf{P}}_{\mathbf{L}}$	84
3.2.2	Statistics of $\tilde{\mathbf{P}} \overline{\mathbf{P}}_{\mathbf{L}}$ under \mathcal{H}_0	84
3.2.3	Statistics of $\tilde{\mathbf{P}} \overline{\mathbf{P}}_{\mathbf{L}}$ under \mathcal{H}_1	85
3.2.4	Mean and variance of $\tilde{\mathbf{P}} \overline{\mathbf{P}}_{\mathbf{L}}$ under \mathcal{H}_0	86
3.3	Statistical tests applied to the standardized periodogram	86
3.3.1	Test of the maximum periodogram value (T_M)	87
3.3.2	Fisher's test and its variations (T_F , T_{Ch} , $T_{F,rob}$)	92
3.3.3	Testing the N_C^{th} largest periodogram value (T_C)	93
3.3.4	Adaptive tests (HC , BJ)	95
3.4	Standardization using parametric estimates of S_E	95
3.4.1	AR parameters estimation	95
3.4.2	Distribution of $\tilde{\mathbf{P}} \hat{\mathbf{S}}_{E,AR}$	97
3.5	Numerical studies on synthetic noise processes	98
3.5.1	Classical tests based on $\tilde{\mathbf{P}} \overline{\mathbf{P}}_{\mathbf{L}}$	99
3.5.2	Adaptive tests based on $\tilde{\mathbf{P}} \overline{\mathbf{P}}_{\mathbf{L}}$	104
3.5.3	Effect of ignoring noise correlations	109
3.5.4	Comparison of the proposed standardization with a parametric approach	111
3.5.5	Application to the design of observational strategies	114
3.6	Summary and conclusions	115

4 Extension to the case of uneven sampling 119

4.1	Effect of uneven sampling on the periodogram	120
4.1.1	Considered sampling patterns	120
4.1.2	Classical periodogram and uneven sampling	123
4.2	Variants of the periodogram	123
4.2.1	The Lomb-Scargle periodogram	123
4.2.2	Least-square fitting and other related periodograms	126
4.3	Generalized Extreme Value distribution	128
4.3.1	Generalities	128

4.3.2	Parameters estimation	129
4.3.3	Diagnostic plots	130
4.4	False alarm rates evaluation	132
4.4.1	Method of “independent” frequencies	132
4.4.2	Bootstrap methods	133
4.4.3	FA evaluation using theory of the extremes	134
4.4.4	Conclusions	135
4.5	Proposed bootstrap procedure	136
4.5.1	Direct bootstrap approach	136
4.5.2	“Accelerated” bootstrap approach	141
4.6	Numerical studies on a synthetic noise process	141
4.6.1	Validation on known processes	141
4.6.2	“GEV accelerated” algorithm	148
4.7	Conclusions	151
	Conclusions and perspectives	153
	Appendices	
	A Theoretical distributions	162
	B Periodogram distribution and non centrality parameter	165
	C Theoretical AR autocorrelation function	170
	D Generalized extreme value : proofs of equations	171
	Bibliography	177

Notations and Definitions

The following lists give the general notation, acronyms, abbreviations, and physical quantities that are frequently used throughout this manuscript. Special notation that is used less frequently will be defined when needed.

General notation

N	Number of data points in the time series
t_j	Observation epoch
$X(t_j)$	Sampled time series
$E(t_j)$	Zero-mean stationary Gaussian colored noise
ν, ν_k	Continuous frequency, Fourier frequency
Ω	Indices set of considered Fourier frequencies
$S_E(\nu)$	Noise power spectral density
r_E	Noise autocorrelation function
$N_S, \alpha_q, f_q, \varphi_q$	Number of sines, sines' amplitude, frequency and phase
N_C	Proxy for N_S
N_p	Number of exoplanets orbiting target star
$T_p, M_p, e_p, \omega_p, t_0, V_0$	Keplerian planet parameters
d_p, R_p	Planet semi major-axis and radius
i	Planet orbital inclination
$a_p(t_j), E_p(t_j), M_{E,p}(t_j)$	True, eccentric and mean anomalies
K_p	RV semiamplitude,
$V_R(t_j)$	Radial velocity
λ, λ_0	Observed and rest wavelength
M_\oplus, M_J	Earth and Jupiter mass
R_\odot, M_\odot	Solar radius and mass
R_\star, M_\star	Stellar radius and mass
$P(\nu)$	Classical periodogram
L	Number of available training data sets
$\bar{P}_L(\nu)$	Periodogram averaged with L training data sets of the noise
$\tilde{P}(\nu \bar{P}_L)$	Periodogram standardized by \bar{P}_L considered at frequency ν
$D_N(\nu)$	Dirichlet kernel
$K_N(\nu)$	Fèjer kernel

Z, z	Scalar random variable, one realization of Z
$v_Z(z)$	$\Pr(Z > z)$ (observed p-value)
V_Z	P-value as a random variable ($V_Z \sim \mathcal{U}_{[0,1]}$)
$\mathbf{Z} = [Z_1, \dots, Z_N]^\top$	Vector of random variables
$V_{\mathbf{Z},k}$	P-value associated to component Z_k of \mathbf{Z}
$V_{\mathbf{Z},(k)}$	k -th ordered p-value of \mathbf{Z}
$\varphi_{\mathbf{Z}}$	Probability density function of \mathbf{Z}
$\Phi_{\mathbf{Z}}$	Cumulative distribution function of \mathbf{Z}
$\mathcal{H}_0, \mathcal{H}_1$	Null and alternative hypothesis
$\text{Bin}(a, b)$	Binomial distribution
$\mathcal{U}_{[a,b]}$	Uniform distribution between a and b
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution of mean μ and variance σ^2
$I_x(a, b)$	Regularized incomplete Beta function
$\lambda_k := \lambda(\nu_k)$	Non centrality parameter
$\rho_k := \rho(\nu_k) := \frac{SE(\nu_k)}{N\sigma_k^2}$	A term relative to the noise PSD (cf. Eq. (B.7))
$\chi_{d_1}^2$	Chi-square distribution with d_1 degree of freedom
$\chi_{\lambda_k}^2(d_1)$	Non central χ^2 distribution with d_1 degree of freedom
$F(d_1, d_2)$	F (Fisher-Snedecor) distribution with d_1 and d_2 degrees of freedom
$F_{\lambda_k}(d_1, d_2)$	Non central Fisher-Snedecor distribution with d_1 and d_2 degrees of freedom
$T_M(Z)$	Statistic of test of the maximum Z value
$T_C(Z)$	Statistic of test of the N_C^{th} maximum Z value
$T_F(Z)$	Statistic of Fisher test
$T_{Ch}(Z)$	Chiu test
$T_{F,rob}(Z)$	Statistic of Robust Fisher test
\mathcal{L}, ℓ	Likelihood and log-Likelihood function
$T_{LR}(Z), T_{GLR}(Z)$	Statistic of Likelihood Ratio and Generalized Likelihood Ratio tests
$T_{VAR}(Z)$	Statistic of the test of the variance
$KS(Z)$	Statistic of Kolmogorov-Smirnov test
$HC(Z)$	Statistic of Higher Criticism test
$BJ(Z)$	Statistic of Berk-Jones test
\Pr	Probability
P_{FA}	False alarm probability
P_{DET}	Detection probability
P_M	Probability of missed detection
γ	Detection threshold
$\hat{S}_{E,AR}(\nu, \hat{\boldsymbol{\theta}}_{AR})$	Noise power spectral density estimated by AR process with $\hat{\boldsymbol{\theta}}_{AR} := [\hat{o}, \hat{c}_j, \hat{\sigma}^2]^\top$ the estimated parameter vector (order, coefficients, and predicted error variance).
P_{LS}	Lomb-Scargle periodogram
$\hat{D}(\hat{P}_{FA})$	Distribution of the false alarm probability estimates
$G(z, \boldsymbol{\theta}_{GEV})$	GEV cumulative distribution function with $\boldsymbol{\theta}_{GEV} := [\xi, \mu, \sigma]^\top$ the GEV parameter vector (shape, location, and scale).
$\mathbf{B}_0, \mathbf{B}^*$	Proposed bootstrap procedure and ‘‘GEV accelerated’’ version

Acronyms

1 D	One-Dimensional
3 D	Three-Dimensional
AR	AutoRegressive (process)
AU	Astronomical Unit
AUC	Area Under Curve
CDF	Cumulative Distribution Function
CFAR	Constant False Alarm Rate
CNES	Centre National d'Études Spatiales
CoROT	COnvection, ROTation et Transits planétaires
ESA	European Spatial Agency
ESPRESSO	Echelle SPectrograph for Rocky Exoplanet and Stable Spectroscopic Observations
FA	False Alarm
FT	Fourier Transform
GEV	Generalized Extreme Value
GLR	Generalized Likelihood Ratio
GOF	Goodness-Of-Fit
GOLF	Global Oscillations at Low Frequency
HARPS	High Accuracy Radial velocity Planet Searcher
HC	Higher Criticism
HD	HydroDynamical (simulations)
HZ	Habitable Zone
IAU	International Astronomical Union
LR	Likelihood Ratio
LS	Lomb-Scargle (periodogram)
MDI	Michelson Doppler Imager
ML(E)	Maximum Likelihood (Estimate)
NASA	National Aeronautics and Space Administration
PDF	Probability Density Function
PLATO	PLAnetary Transits and Oscillations of stars
PSD	Power Spectral Density
ROC	Receiver Operating Characteristic
RV	Radial Velocity
SNR	Signal-to-Noise Ratio
SoHO	Solar and Heliospheric Observatory
WGN	White Gaussian Noise

Abbreviations

cf.	<i>confer</i> (refer to)
e.g.	<i>exempli gratia</i> (for the sake of example)
i.e.	<i>id est</i> (that is)
i.i.d.	independent and identically distributed
inf	least lower bound
resp.	respectively
r.v.	random variable
sup	greatest upper bound
<i>vs</i>	<i>versus</i>
w.r.t.	with respect to

Physical quantities

$c = 299\,792\,458 \text{ m}\cdot\text{s}^{-1}$	Speed of light
$G = 6.67408 \times 10^{-11} \text{ m}^3\cdot\text{kg}^{-1}\cdot\text{s}^{-2}$	Gravitational constant

General Overview

Introduction

Since the detection of the first extrasolar planets in the 90's, several thousands of new worlds have been discovered. As the direct detection of planets is hard, the majority of the detection techniques are indirect. For example, the radial velocity (RV) technique consists in measuring the stellar motion through wavelength shifts (Doppler effect) due to the gravitational attraction of orbiting planets. Modern instrumental performances have reached signal-to-noise ratios allowing, in principle, to detect Earth mass planets by the RV method. However, at such low levels of instrumental noise, a new critical and limiting issue appears. The stellar surface can be seen as a boiling fluid, with millions of convection cells generating upward and downward plasma flows visible under the form of granules. These motions generate random fluctuations in the measured RV of the star that can mimic, or hide, exoplanetary signatures. The statistical characterization of this stellar background noise is difficult, and this specific issue motivated the present study.

Assessing the significance of a detection claim when the noise statistics are fully or even partially unknown is a challenging problem. A fundamental quantity/tool when searching for frequencies in temporal time series is the confidence level. This measure of the detection significance is related to the statistical distribution of the noise within the data. Knowing the noise statistics allows to derive detection tests of which we control the statistics and can evaluate the false alarm probability. In practice, applying test statistics without considering these noise correlations can lead to uncontrolled false alarm rate. This is especially important when the signal-to-noise ratio is low, i.e. when the planet signature magnitude is at the level of the jitter induced by the convective motions, as for telluric planets orbiting variable stars. Statistical tests and analysis accounting for spectral estimation are then essential to derive such confidence intervals and they will take a significant part of this manuscript.

Motivation

A key point is that in parallel to technological advances, astrophysicists have continuously improved stellar models and elaborated numerical simulation codes able to account for the complex interplay of various astrophysical processes in the star's interior and surface. Recent works demonstrate that granulation noise can be reproduced in a reliable way by large scale hydrodynamic (HD) numerical simulations. This suggests the possibility of using such *ab initio* simulations in the detection process.

My thesis aims to design and analyze reliable tools to control the false alarm and detection probabilities when the noise is colored with unknown statistics. To deal with partially unknown statistics, we investigate, through statistical methods, whether *ab initio* HD simulations could improve the control of the significance assigned to planet detections. In other words, we

consider that the noise statistics are partially known through independent simulations.

The goal of these works is to analyze, for the first time, the probability of false alarm and detection of various tests conducted in presence of stellar convective noise. It is connected to different fields of research including signal processing, planet detection and stellar physics.

Contributions

Chapter 1 introduces the astrophysical context of the thesis. The main contributions follow in the next chapters.

- Chapter 2: This chapter gives a review of signal detection theory. An analyze of the Keplerian signatures in the temporal and Fourier domains is given.
- Chapter 3: This chapter investigates whether independent simulations of the colored noise could improve the detection significance. Using simulated time series to calibrate the periodogram of the observations, we design a frequency analysis tool that takes the form of a *standardized periodogram*. In the case of regularly sampled observations, the statistical analysis of such periodogram demonstrates that it is asymptotically independent of the colored noise. This important property allows to design several detection tests that turn out to be Constant False Alarm Rate detectors. For some of these tests, we derive analytical expressions of the false alarms and detection probabilities. Among the investigated tests, particularly attention is given to tests adaptive in the number of periodic components. Such tests, as the Higher-Criticism or the Berk-Jones tests, were recently largely investigated in the Signal Processing community. Never used in the exoplanet detection community, we propose to apply them to the periodogram. We found that these adaptive tests appear more powerful than classical tests based on the highest periodogram component in case of multi-signals, off-grid planetary period, and high eccentric orbits. The performed statistical analysis allows the design of observational strategies for several detection tests.
- Chapter 4: Considering the test based on the maximum periodogram value, we investigate in this chapter the control of the test statistics in the case of irregularly sampled observations and colored noise. Because the distribution of any periodograms is not known when the noise is not white and the sampling not regular, we develop a bootstrap procedure adapted to the proposed standardized periodogram. By generating thousands of false alarm probability estimates, this bootstrap provides access to the distribution of these estimates in order to bound the true false alarm probability, independently of the data sampling. To save a significant amount of computational time, we further show that the procedure can be improved by using generalized extreme value distributions. The technique is an extension of previous works developed for the white noise case (cf. [Süveges, 2014]) to the case of partially unknown colored noise.
- Works under investigation: In perspective of this thesis, we present a preliminary application of the proposed methodology to real astrophysical data. We show a preliminary comparison of radial velocities extracted from recent 3D hydrodynamic simulations of the stellar convection with real solar data taken from GOLF spectrophotometer onboard the SoHO spacecraft. Then, we discuss further other applications of this thesis work.

Several scientific publications were presented during this thesis. The following list refers to the refereed publications.

International refereed journals:

1. S. Sulis, D. Mary, and L. Bigot. A study of periodograms standardized using training datasets and application to exoplanet detection. *IEEE Transactions on Signal Processing*, 65(8):2136-2150, April 2017. 10.1109/TSP.2017.2652391.
2. **[In preparation.]** S. Sulis, L. Bigot, and D. Mary. Using hydrodynamical simulations of the stellar convective noise to control exoplanet detection. i) the case of regularly sampled observations.

Referred publications as a book chapter of a Summer School:

- 3 S. Sulis, D. Mary, and L. Bigot. Overcoming the stellar noise barrier for the detection of telluric exoplanets: an approach based on hydrodynamical simulations. In D. Mary, R. Flamary, C. Theys, and C. Aime, editors, *EAS Publications Series*, volume 78-79 of *EAS Publications Series*, pages 247-274, September 2016. 10.1051/eas/1678012.

Referred publications in the proceedings of international conferences:

- 4 S. Sulis, D. Mary, and L. Bigot. A bootstrap method for sinusoid detection in colored noise and uneven sampling. Application to exoplanet detection. In *25th European Signal Processing Conference (EUSIPCO)*, pages 1095-1099, August 2017. 10.23919/EUSIPCO.2017.8081377
- 5 S. Sulis, D. Mary, and L. Bigot. Using hydrodynamical simulations of stellar atmospheres for periodogram standardization: Application to exoplanet detection. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4428-4432, March 2016. 10.1109/ICASSP.2016.7472514.

Other referred proceedings (in french):

- 6 S. Sulis, D. Mary, and L. Bigot. Sur l'utilisation de rapports de périodogrammes pour la détection de raies dans un bruit de densité spectrale partiellement inconnue. In *XXVème colloque GRETSI*, pages 4428-4432, September 2015.

Organization of the dissertation

This manuscript is divided in two main parts, and a review of the works under investigation is given in conclusion.

- Part I introduces the context of extrasolar planet detection and some fundamentals in detection theory as well as several tools related to this thesis framework (hypothesis testing, periodogram, tests statistics). Chapter 1 gives a survey of the exoplanet detection field from the 90's up to the present days, and describes the radial velocity technique as well as the different noise sources we have to face in order to be able to detect Earth-mass planets orbiting variable stars. Chapter 2 deals with the problem of detecting periodic signatures in noisy data. We look at the particular cases of different kind of noises (white and colored) and at the planetary signatures in the temporal and Fourier domains.

- Part II presents the main contributions of this research work. Chapter 3 introduces a new tool, namely a specific type of standardized periodogram, and make a statistical analysis of different detection tests in the case of a regular time sampling. The false alarm and detection probabilities are analytically derived for some of these tests, and some detectors are proved to be CFAR. The advantages of using adaptive tests designed to detect non extremely sparse signatures are also investigated for detecting planetary signatures. Chapter 4 propose an adapted bootstrap procedure based on such periodograms to control the detection significance in case of irregular sampling. In a second step, Generalized extreme values are exploited to save a large amount of computational time.
- Finally, in Perspectives, we present a preliminary application of this work to real astrophysical data. We investigate the realism of hydrodynamical simulations of the solar convective noise with real observations of the solar convection. Then, we discuss several applications of the developed study regarding the extrasolar solar planet detection field.

List of Figures

1.1	Illustration of the principle of the radial velocity technique.	7
1.2	Properties of the detected extrasolar planets depending on the detection method.	8
1.3	Planet mass <i>vs</i> date of discovery.	9
1.4	RV curve of 51 Pegasi b, the Earth, Jupiter, and the entire Solar System planets.	11
1.5	Illustration of the three anomalies associated to an elliptical orbit.	12
1.6	Effect of the orbit orientation on an Earth-like planet RV signature	13
1.7	RV accuracy depending on the spectral type.	14
1.8	RV semi-amplitude <i>vs</i> planet period.	16
1.9	Illustration of the solar colored noise.	18
1.10	Image of the solar disk in visible light taken by the instrument MDI/SoHO.	19
1.11	Illustration of the distortion of the spectral line shape due to starspot.	20
1.12	Emergent intensity at the solar surface: observation <i>vs</i> simulation.	22
1.13	Dopplergrams of the Sun taken by the MDI/SoHO.	24
1.14	Blueprint of a local box simulation and 3D snapshot with the velocity field.	26
1.15	Power spectral density estimate of simulated solar RV at three different epochs.	27
1.16	Illustration of the granulation at the stellar surfaces.	28
2.1	Influence of the parameters of a central and non F central distributions.	35
2.2	Test distribution under both hypotheses and decision regions plot.	35
2.3	Illustration of the P -value for a test statistic F-distributed.	37
2.4	Illustration of a test performance through the ROC curve and AUC.	39
2.5	Dirichlet kernel and spectral leakage effect on a sinusoidal signal.	43
2.6	Periodogram distributions at one Fourier frequency under both hypotheses.	48
2.7	Verification of the asymptotical bias and variance of the periodogram.	51
2.8	Influence of the Keplerian parameters on RV planet signatures.	53
2.9	RV planet signatures in the Fourier domain depending on the Keplerian parameters.	55
2.10	Histograms of multi-planetary system and orbital eccentricities.	56
2.11	Comparison of the performance of T_M and T_{VAR}	61
2.12	ROC curves of tests T_F , T_{Ch} , and $T_{F,rob}$ for $N_S = 10$ sinusoids in noise.	65
2.13	Illustration of the principle of the HC test statistics and Detectability region	68
2.14	Illustration of HC test statistics applied to $N_S = 40$ sines in a WGN.	70
3.1	Mean and variance of the standardized periodogram.	87
3.2	Density of $T_M(\tilde{\mathbf{P}} \bar{\mathbf{P}}_L)$ for various values of N and L	91
3.3	Theoretical mode, esperance and variance of $T_M(\tilde{\mathbf{P}} \bar{\mathbf{P}}_L)$	92
3.4	Distribution of the AR order estimates for various criteria.	97
3.5	Illustration of the considered AR(6) process.	99
3.6	Colored noise: theoretical <i>vs</i> empirical results for $T_M(\tilde{\mathbf{P}} \bar{\mathbf{P}}_L)$ and $T_C(\tilde{\mathbf{P}} \bar{\mathbf{P}}_L)$	100
3.7	ROC curves of tests $T_M(\tilde{\mathbf{P}} \bar{\mathbf{P}}_L)$ and $T_C(\tilde{\mathbf{P}} \bar{\mathbf{P}}_L)$ showing the influence of L and N_C	102
3.8	Influence of the time series length on $T_M(\tilde{\mathbf{P}} \bar{\mathbf{P}}_L)$ performance.	102
3.9	Differences between theory and empirical evaluation of the P_{FA} and P_{DET}	103

3.10	Illustration of different order of correlations on the theoretical results.	104
3.11	Adaptive tests HC and BJ <i>vs</i> classical tests T_M and T_C applied to $\tilde{\mathbf{P}} \overline{\mathbf{P}}_L$	106
3.12	Study under \mathcal{H}_0 of tests T_M , HC and BJ applied to $\tilde{\mathbf{P}} \overline{\mathbf{P}}_{20}$	108
3.13	Probability of detecting the correct support under \mathcal{H}_1 for T_M , HC, and BJ.	108
3.14	Comparison with methods ignoring noise correlations.	110
3.15	Histograms of the frequency distribution under \mathcal{H}_0 and \mathcal{H}_1 for different tests.	110
3.16	Comparison with parametric approaches: ROC curves.	112
3.17	Parametric approaches: $P_{FA}(\gamma)$	113
3.18	Parametric approaches: $P_{FA}(\gamma)$ for different L	114
3.19	Example of a detectability study for a planet orbiting a solar type star	115
4.1	Sine function for different sampling patterns.	122
4.2	Irregular sampling: classical periodogram <i>vs</i> LS periodogram.	125
4.3	Diagnostic plots for three different GEV estimations.	132
4.4	P_{FA} estimation procedure (possible only if $L \gg 1$).	137
4.5	P_{FA} estimation procedure if $1 < L \ll \infty$	138
4.6	Blueprint of the B_0 procedure.	140
4.7	Snapshots of some steps in the B_0 procedure for an AR(6) colored noise process.	142
4.8	B_0 procedure: \hat{P}_{FA} <i>vs</i> P_{FA}	143
4.9	P_{FA} estimation using a classical bootstrap procedure ignoring the noise correlations.	143
4.10	B_0 procedure: Pdf of T_M for the three samplings.	145
4.11	B_0 procedure: Repartition in frequency of largest periodogram peak.	145
4.12	B_0 procedure: $\hat{P}_{FA}(\gamma)$ <i>vs</i> true $P_{FA}(\gamma)$	146
4.13	$\hat{\mathcal{D}}_{\hat{P}_{FA}}(B_0)$ for different samplings.	146
4.14	$\hat{\mathcal{D}}_{\hat{P}_{FA}}(B_0)$ for 10 realizations with different random irregular samplings \mathcal{S}_3	147
4.15	B_0 <i>vs</i> B^* procedures: $\hat{P}_{FA}(\gamma)$ comparison.	148
4.16	B_0 <i>vs</i> B^* procedures: $\hat{\mathcal{D}}_{\hat{P}_{FA}}$ comparison.	149
4.17	B_0 <i>vs</i> B^* procedures: dispersion, mean and, computational time comparisons.	150
4.5	Spectrum of the solar Sodium Doublet and blueprint of the RV extraction process.	156
4.6	Comparison of solar RV data with 3D HD simulations.	157

List of Tables

1.1	Orbital characteristics and RV semi-amplitudes of the Solar System planets and some extrasolar planets.	11
1.2	Properties of the solar granulation and supergranulation.	23
2.1	Probabilities involved when evaluating a decision rule.	32
2.2	Mode, mean, and variance of the central F law	33
2.3	Mode, mean, and variance of the non-central F law	34
2.4	Numerical difference in terms of P_{FA} and P -value measured on the F and Gaussian distributions.	37
2.5	Table of the some test statistics found in the literature for white noise	76
A.1	Some properties of the Gaussian distribution.	162
A.2	Some properties of the χ^2 law.	163
A.3	Some properties of the non central χ^2 law.	163
A.4	Some properties of the central F law.	164
A.5	Some properties of the non-central F law.	164

Part I

Introduction

Chapter 1

Exoplanet detection

Contents

1.1	Methods and discoveries	4
1.2	The radial-velocity technique	10
1.2.1	Planetary signatures	10
1.2.2	Method of measurement and instrumental errors	13
1.3	Stellar noises	16
1.3.1	Magnetic activity	17
1.3.2	Oscillations	20
1.3.3	Convection	21
1.3.4	Summary	25
1.4	Simulations of the convective stellar activity	25
1.5	Conclusions on the astrophysical context of the thesis	27

Science, and especially astrophysics, is a dynamic field evolving very fast. Our understanding about the Nature’s laws and about our place in the Universe is always progressing and new informations force ourselves to continuously question our basic theories and definitions. The formal definition of a *planet*, for example, was discussed during a long time after the discovery in 2005 of thousands of transneptunian objects¹ in the Solar System, as Éris, a Pluto size object [Soter, 2006]. It was finally decided, in 2006, by the *International Astronomical Union* (IAU) that the formal word “planet” concerns only the eight Solar System planets.

Definition 1. (Solar System) planet

A planet is defined by the Resolution B5 of the IAU² (2006) as a celestial body that (a) orbits around the Sun, (b) has sufficient mass for its gravity to overcome rigid body forces and assume hydrostatic equilibrium (a nearly round shape), and (c) has cleared the neighbourhood around its orbit.

In the Solar System, we found rocky planets close to the Sun (Mercury, Venus, the Earth and Mars), gaseous giant planets beyond the ice line (Jupiter, Saturn, Uranus and Neptune) and dwarf planets beyond (e.g. Pluton, Éris). However, this definition is difficult to apply to bodies detected around other stars than the Sun. Indeed, in practice, no information about the detected object shape is available (condition b) as well as the environment in which this object evolves (condition c). Officially, the IAU has not yet decided on a definitive definition of the word *extrasolar* planet.

¹i.e. objects orbiting behind the Neptune’s orbit.

²https://www.iau.org/static/resolutions/Resolution_GA26-5-6.pdf

Definition 2. Extrasolar planet

The “Working Group on Extrasolar Planets” proposed³, in 2003, to define an extrasolar planet as (a) a body orbiting an other star than the Sun (or a stellar remnant), (b) with a mass below the mass limit for thermonuclear fusion of deuterium ($\approx 13 M_J$ for objects of solar metallicity), and (c) with a minimum mass and/or size defined as for our Solar System planets. Substellar objects with masses above the limiting mass for thermonuclear fusion of deuterium are “brown dwarf”. Free-floating objects (i.e. planetary bodies not gravitationally linked to a star), in young star clusters, with masses below this mass limit are sub-brown dwarfs.

However, this “working definition” is debated and alternative ones are currently elaborated, which are related to the object characteristics, circumstances or formation history (cf. [Basri and Brown, 2006] for an interesting discussion on the subject). In this study, it will be sufficient to consider extrasolar planets as celestial bodies orbiting other stars than the Sun (and that are not double stars).

For more than 20 years, we know that a multitude of other worlds exist. How do we detect them ? Are we able to detect Earth-like planets⁴ ? The objective of this chapter is to give a short (necessary not complete) introduction of the astrophysical context underlying these questions. In a first part, we propose a survey of this topic through the different methods of planet detection and a recent review of the main discoveries (Sec. 1.1). We will then introduce the particular technique of the radial velocities and discuss the main limitations to detect Earth-mass planets through this method (Sec. 1.2). We will see that the instrumental performance have been improved a lot during the last two decades, so that nowadays the main limitation to detect the smallest extrasolar planets comes from the “noise” of the intrinsic activity of the host star itself (Sec. 1.3). We will finally present the hydrodynamical (HD) simulations of the stellar surface convection, which is a stochastic contribution of the stellar noise (Sec. 1.4), and conclude on the thesis issue (Sec. 1.5).

1.1 Methods and discoveries

The quest of extrasolar planets began in the 90’s, with the discovery of several planets orbiting around a pulsar named PSR 1257+12 [Wolszczan and Frail, 1992]. This type of star is a rapidly rotating neutron star with an extremely dense core resulting of the explosion of a massive supernova star. The astronomers were greatly surprised to find that planets have survived to the stellar explosion.

Few years later, the first extrasolar planet orbiting a Sun-like star was discovered: 51 Pegasi b [Mayor and Queloz, 1995]. Once again, this detection surprised the astronomers since the planet is a giant (around half of the Jupiter mass) orbiting very close to the host star with a very short period of only 4.2 days. This type of planets, now named *Hot Jupiter* planets⁵, are not met in the Solar System. This discovery turned the old planetary formation models⁶ upside-down and new theories of systems formation based on planet’s migrations were then developed (e.g. [Nelson et al., 2000, Masset and Papaloizou, 2003, Morbidelli et al., 2005, Tsiganis et al., 2005]). This detection was done by the *radial velocity* technique (described in the items below) and announced a new era in astrophysics: the hunt of extrasolar planets.

Since then, many space photometric missions (e.g. COROT [Baglin et al., 2006], Kepler

³<http://w.astro.berkeley.edu/~basri/defineplanet/IAU-WGExSP.htm>

⁴i.e. Earth-mass, density and temperature planets.

⁵Hot Jupiters are defined as Jupiter-type planet orbiting at < 0.1 AU.

⁶Standard models explained the planetary formation by a core accretion effect beside the snow line (i.e. at large distance from star).

[Borucki et al., 2010]) and ground based spectroscopy surveys (e.g. ELODIE, SOPHIE [Bouchy, 2015], HARPS [Mayor et al., 2003]) have been dedicated to exoplanet hunting. A multiplicity of new detection methods have emerged and the instrumental precision improved to end up around 3610 extrasolar planets detections (June, 2017, source: exoplanets.eu). Following [Perryman, 2011], the exoplanet detection techniques can be classified in four main families:

Direct observation: The direct imaging method consists in taking a snapshot of the planets and disk around a central star. While this is the most obvious method, it is also one of the most difficult due to the strong contrast in luminosity between the star and the planet. The method consists in hiding the stellar disk with a stellar mask (i.e. a coronagraph) to be able to detect the stellar light reflected by the orbiting planet, and to correct the image using adaptive optics in real time. In June 2017, this method counted 63 detections. The majority of these planets orbits around young stars or nearby stars [Perryman, 2011]. The success of this method depends on the separation between the star and the planet. When coupled with spectroscopy, it informs on the planet radiation. This allows to constrain, among other, the planet temperature, pressure, chemical composition and the atmosphere structure. For an exhaustive review of this technique, one can report to [Beuzit et al., 2006].

Effects of the host star: Planets create a gravitational force on their star making the system star-planets moving around a common centre of mass. This effect can be detected by measuring periodic (or quasi-periodic) perturbations of the stellar motion projected on the line-of-sight. Depending on the system's inclination and the stellar properties, three different methods exist to detect this motion.

- The radial velocity method (RV), also called the *Doppler spectroscopy technique*, consists in detecting the Doppler shifts in the stellar spectral lines. This shift is related to the stellar motion by the Doppler effect (see Fig. 1.1 and Sec. 1.2). This tiny relative shift (about $\frac{\Delta\lambda}{\lambda} \approx \frac{v}{c} \approx \frac{0.1}{3 \times 10^8} \approx 3 \times 10^{-10}$ for an Earth-like planet), observed as a function of time, depends on the system inclination (systems seen pole-on, i.e. $i = 0^\circ$, cannot be detected by this method). To date, 626 planets have been detected using the RV technique, mostly around FGK stars as they present a large number (several thousands) of spectral lines making the detection of the tiny wavelength shift easier.
- The astrometry method consists in detecting periodic variations (wobbles) of the stellar angular position in the sky as function of time. This method needs a very high precision on the star location and, to date, there are only very few planets detected by this method (in fact, only 1 is listed in the exoplanets.eu catalogue [Muterspaugh et al., 2010]). But far more detections are expected in the next years with the results of the GAIA satellite (2014-2022) [Perryman et al., 2001, Casertano et al., 2008]. According to Perryman et al., around 50 000 – 90 000 new detections are expected after 10 years of this space mission (mostly giant planets with relatively long period, i.e. < 5 AU) [Perryman et al., 2014]. If such an amazing estimation is correct, this will make the astrometry technique the most efficient method, in terms of number of detected planets, for the next decades.
- The timing method, applied to pulsar stars, consists in detecting variations in the time of arrival of the stellar radiative electromagnetic radiation. Indeed, the timing of this pulse is extremely regular and can be compared to the timing of atomic clocks. Small variations can indicate the pulsar motion due to orbiting masses. The timing depends on the star's rotation and can take values of some milliseconds to seconds. In June 2017, 23 planets have been detected around pulsars (among which we find

the three planets orbiting PSR 1257+12, introduced as the first exoplanets detection some lines above). The few cases of such pulsar’s planets seem to indicate that the first impression of the astronomer, questioning the possibility of their existence after a supernova explosion, was right and that planets orbiting pulsar are rare.

These three methods, based on the gravitational effect of planets on their host stars, allow to derive the planet mass and period. One can respectively report to [Fischer et al., 2014], [Perryman et al., 2001] and [Schneider, 2005] for recent reviews of these techniques.

Photometric effects: When a planet crosses the stellar disk on the line-of-sight of the telescope, one can observe a tiny decrease of the stellar flux as the planet blocks a part of the stellar light (e.g. a flux decrease of 80 parts per million is expected for an Earth-size planet orbiting a Sun-like star). This technique, called the *transit method* is, nowadays, the most efficient for detecting planets, with about 1272 detections and more than 3000 planetary candidates. While the probability of observing a planetary transit is small⁷, this huge number of discoveries was possible thanks to dedicated space missions as COROT (led by CNES), a mission that operated from 2006 to 2014 [Baglin et al., 2006], and Kepler, a NASA mission operating from 2009 up to 2014 [Borucki et al., 2010]. The observation of a planet transit gives informations on the planet radius, distance, and orbital inclination. When transmission spectroscopy is done during the planetary transit, it informs on the chemical composition and temperature of the planet atmosphere (e.g. study of HD 209458 [Brogi et al., 2017]). Infrared observations during the secondary transit (i.e. when planet is crossing behind the star) allowed to make the first planetary spectra and to detect molecules in the atmospheres (e.g. atomic sodium found in the atmosphere of HD 209458 b [Deming et al., 2013]). One can report to [Seager, 2008] for an exhaustive review of this technique.

Gravitational microlensing effects: The concept of gravitational microlensing is due to Einstein, who predicted that the distortion of spacetime by a massive object produces an observable light magnification due to the gravitational field [Einstein, 1936]. Discovering planets by the *microlensing* method consists in detecting the bending of light of a distant bright source (*background star*) when it is aligned with a massive compact system containing one or more planets (*lens system*) (e.g. [Gould and Loeb, 1992]). Microlensing events are rare. Since the last decades, a large ground based telescopes network has been dedicated to this technique on the three South hemisphere continents (MOA, OGLE, LCOGT, KMTNet, PLANET et μ FUN [Beaulieu et al., 2013]) in order to maximize the possibility of observing such events. This network works with alerts and allows to obtain continuous measurements with high cadence during a detected event. The typical duration of an event is about 25 days (e.g. [Ranc et al., 2015]). If the observational conditions are optimal (i.e. a good temporal coverage of the event: no data gap, a good cadence: typically 20 minutes during the totality of the event, and of course, a good precision), refined models can be used to describe the event. Simple microlensing models allow to measure the mass ratio between the planet and the host star as well as the projected separation between them (known up to a factor called the *Einstein beam*). By measuring the microlensing parallaxes, with satellite such as Spitzer, K2, and (in the future) WFIRST other properties of the star/planets system can be derived (e.g. the total mass of the system or

⁷The probability that a planet’s orbit, with a random orientation, transits the star is proportional to the planet semi major axis (d_p) and the stellar radius (R_*) [Borucki and Summers, 1984]

$$\Pr(\text{“transit”}) \approx \frac{R_*}{d_p}.$$

For example, the probability of detecting a Jupiter planet, at 5 AU, orbiting a Sun is about 0.1% and the probability of detecting a Hot-Jupiter located at 0.05 AU is about 10%.

the Earth-star distance). In June 2017, 43 planets have been detected by microlensing. This technique is actually the most efficient to detect planets at large distances but the confirmation of such detections by other techniques is also the most difficult. One can refer to [Bennett, 2008, Beaulieu et al., 2006, Cassan, 2014, Ranc et al., 2015] for recent reviews of this technique.

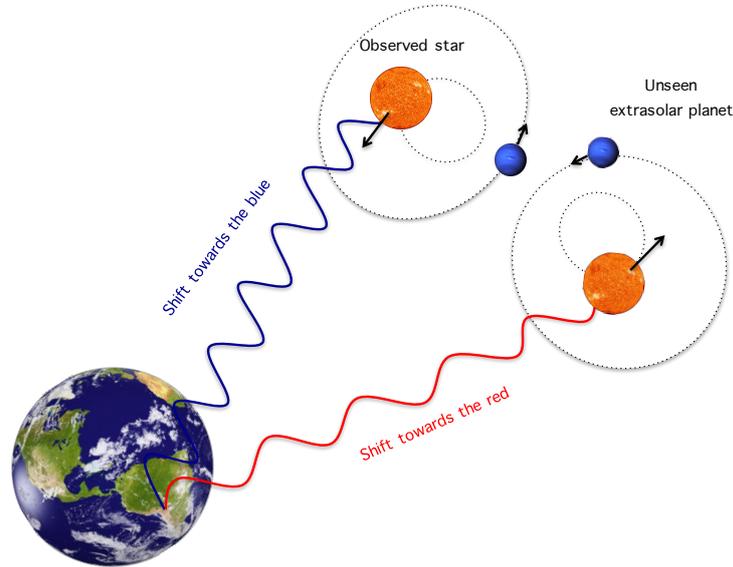


Figure 1.1 – Illustration of the principle of the radial velocity technique. When the star moves toward the observer, it produces a Doppler shift of the stellar lines toward the blue and toward the red when it moves away. This periodic shift allows to detect indirectly the presence of planets orbiting the observed star and gives a lower limit of the planet mass.

When the detection of a celestial body by two of these different methods is possible, a detection is officially claimed as *confirmed*. Moreover, the combination of two of them can allow the characterization of the planet properties. For example, the combination of the RV (mass) and transit (radius) informations gives an estimate of the average planet density, which may be an indicator of the liquid water presence on the planet surface. This precious information, combined with the star-planet parameters, allows to draw preliminary models for the planet structure. Among these models, one can find: telluric planets (e.g. Earth mass planets [Shvartzvald et al., 2017]), Super-Earth⁸ [Santos et al., 2004], ocean or carbon planets [Léger et al., 2004, Kuchner and Seager, 2005], Hot-Jupiters [Mayor and Queloz, 1995]), chthonian planets⁹ [Hébrard et al., 2004], or giant planets (e.g. mini-Neptune [Barnes et al., 2009]).

Fig. 1.2 illustrates the discovered planets as a function of their masses and their distances to the host star. The color code indicates the different detection methods. These discoveries reveal an extreme diversity of the planet populations in terms of mass and orbital distances. A global observation of this figure indicates that planets were found far much closer to their star than our Solar System planets (as for instance the Hot Jupiter population). This is of course an observational bias as we detect mostly the planet producing the largest signals (observational limit). A new planet population appears to be the most abundant since the last survey of the

⁸Super-Earth are planets with mass between 2 and 10 Earth masses.

⁹Chthonian planets are hypothetical planets, born as giant gas planets with atmosphere composed of hydrogen and helium that escaped with time due to their host star proximity. According to the giant planet interior models, it should remain a solid or metallic rock core [Guillot, 1999].

Kepler satellite [Borucki et al., 2010]: the Super-Earth planets, which are not present in our Solar System.

Today, we know that multi-planetary systems orbiting other stars exist elsewhere (e.g. Kepler-11 hosting a procession of 6 planets [Lissauer et al., 2011] or TRAPPIST-1 hosting a system of 7 planets [Gillon et al., 2017]). Some of these planets orbit their star in less than 1 day (e.g. KOI 1843.03 with an orbital period of 4.2 hours [Rappaport et al., 2013]) and others orbit in the Habitable Zone¹⁰ (HZ) of their star (e.g. Gliese 667c with 3 planets in the HZ [Anglada-Escudé et al., 2013]).

Over the last few years, the number of detections of small telluric planets located in the HZ of their star is increasing very rapidly. Indeed, last year (2016), astronomers discovered a planet orbiting one of our nearest neighbouring star, at only 4 light year: Proxima Centauri, a red dwarf M star [Anglada-Escudé et al., 2016]. The planet has a mass comparable to Earth and orbits into the star's HZ in 11 days (M dwarves are interesting targets as their HZ is far most closer than for Sun-like stars). In the same year, the NASA discovered 7 Earth-size planets orbiting also a M star named Trappist-1 [Gillon et al., 2017].

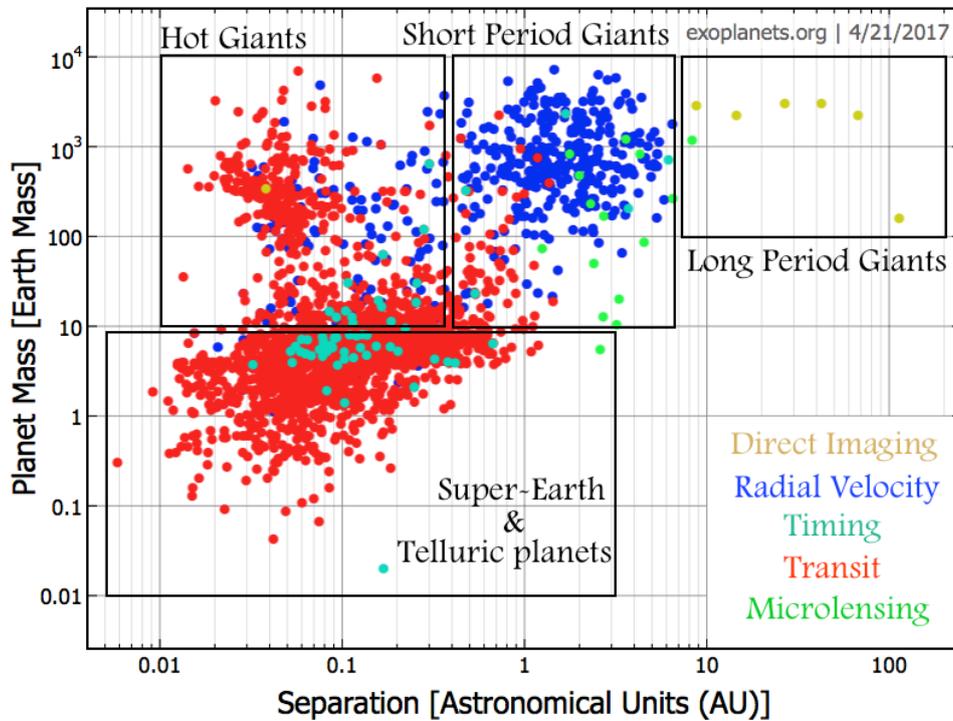


Figure 1.2 – Properties of the detected extrasolar planets depending on the detection method. This figure represents the planets as a function of their masses and distances to their host star. Four categories of planets are highlighted by the different frames. Note, however, that for the planet population that has been detected by transit (red dots), one cannot access to the mass. Consequently, some of the planets present in this graph are either confirmed by an independent method (RV) given access to the mass or the mass has been roughly estimated through a model linking the transit observations with the body's mass. Masses derived by such method are subject to caution. Source: exoplanets.org.

¹⁰According to [Seager, 2013], the habitable zone is the area where liquid water can be found on a planet surface. In our Solar System this zone is located between 0.5 and 3 AU. Of course, the presence of liquid water depends also on other parameters like the planet history or the stellar activity. For example, in our Solar System, Venus, Mars, the Moon and a lot of asteroids are included in this habitable zone whereas no liquid water is present today on their surface. This HZ must just be seen as a preliminary indicator for the search of life such as we know it (i.e. developed in a water ocean at a planet's surface), but it is not excluded that other forms of life can exist in the Universe (e.g. below the surface of some giant planets icy moons).

All of these discoveries result from the exponential improvement of the instrumental performance (e.g. optics, detectors precision and space instruments), the calibration method, and the detection algorithms. Indeed, detecting planets is always a difficult task for all the existing detection techniques as they are far much fainter, less massive, and smaller than their host star. These reasons make the massive giant planets easier to detect in general. This is illustrated by Fig. 1.3 showing the planets discovered by the RV method as a function of their mass and the year of their discovery. We see the increase of the number of small planets detection with time. Between 1995 and 2003, the majority of the detections were done using a small number of instruments observing in the optical domain: HIRES (Hawaii), CORALIE (La Silla, Chili), The Hamilton Spectrograph (Lick), ELODIE (Haute-Provence Observatory, France), AFOE (Whipple Observatory), UCLES (Australia), the Coudé Echelle Spectrograph, the Sandiford Cassegrain Echelle spectrograph and HRS (McDonald Observatory, USA). After 2003, the number of detections drastically increased with the start of the HARPS commissioning (La Silla, Chile) opening the window to the Super-Earth and mini-Neptune planets populations [Mayor et al., 2003]. With the soon coming instrument ESPRESSO (installed on the VLT), the instrumental precision limit will reach the level allowing to detect Earth-like planets orbiting in the HZ of Sun-like stars with the RV method (cf. Sec. 1.2.2) [Pepe et al., 2010].

In the next decade, other important instruments dedicated to the detection of low-mass planets in the HZ of M dwarfs (and observing this time in the near infrared), will be set up: CARMENES, installed in 2016 in the Calar Alto Observatory [Quirrenbach et al., 2014], SPIRou, which will be installed on the Canada-France-Hawaii telescope this year (2017) [Thibault et al., 2012], or still the ExTrA project, which will start in 2017 at La Silla, Chile [Bonfils et al., 2015].

With these new instruments, the hunt of Earth-like planet is became possible.

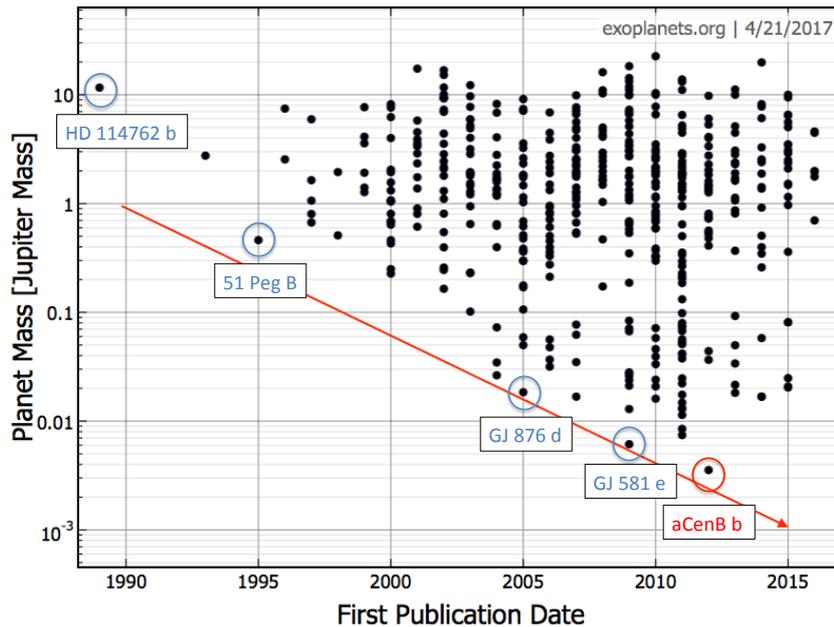


Figure 1.3 – Planet mass *vs* date of discovery. One can see the impressive increase of the number of detections with time up to a factor 1000 in less than 2 decades (1995-2015). Some famous extrasolar planet names are indicated. The name of the planet α Centauri B b, claimed discovered in 2012 in [Dumusque et al., 2012] can be seen in red as this detection is debated. Indeed, other studies have shown that the claimed planet detection can be due to the contribution of the host star surface activity [Hatzes, 2013, Rajpaul et al., 2016]. The different sources of this stellar activity will be detailed in Sec. 1.3. Source: exoplanets.org.

According to recent statistical studies, 16% of the FGK stars of our galaxy have at least one planet [Fressin et al., 2013], and 15% of solar type stars would have an Earth-size planet [Petigura et al., 2013]. Regarding the number of detected planets¹¹, only 35 planets into more than 3600 have a mass $< 2M_{\oplus}$. Assuming the validity of such statistical studies about the existence of such small exoplanets, we can deduce that we are not able to detect them yet (or that their detections are very difficult and hazardous).

1.2 The radial-velocity technique

The radial velocity technique is based on the Doppler-Fizeau effect introduced by Doppler in 1842 [Doppler, 1842]. This effect links the stellar motion (due to the influence of an orbiting planet) on the light-of-sight of the observer to a wavelength shift in the stellar spectra. When the star is going toward the observer, the stellar lines are blueshifted and redshifted when the star is moving away (cf. Fig. 1.3). By ignoring relativistic effects (as $V_R \ll c$), the relationship between the RV and the wavelength shift is:

$$\frac{\lambda - \lambda_0}{\lambda_0} = \frac{V_R}{c}, \quad (1.1)$$

with c the speed of light, λ_0 the wavelength in the reference frame, λ the observed wavelength and V_R the projected velocity (i.e. the radial velocity).

The shape of a RV planet signature depends on the planet orbital parameters. The motion is described by the Keplerian parameters of an elliptical planet orbit. These signatures are described in the following section, which is largely inspired by [Perryman, 2011].

1.2.1 Planetary signatures

The RV semi-amplitude (K) due to a planet orbiting a star is related to the stellar and planetary characteristics by:

$$K = \left(\frac{2\pi G}{T} \right)^{1/3} \frac{M \sin(i)}{(M_{\star} + M)^{2/3}} \frac{1}{\sqrt{1 - e^2}} \quad (1.2)$$

with G the gravitational constant, T the planet period, M the planet mass, M_{\star} the stellar mass, i the inclination of the planet orbit with respect to the reference frame (cf. Fig. 1.6) and e the orbit eccentricity¹² (for the detailed calculations, see p.12 of [Perryman, 2011]). According to (1.2), for a given stellar mass, it is easier to detect massive planets, short period planets, and planets with small eccentric orbit. These remarks are indeed observed in Fig. 1.2, where the detected planets are mostly massive and with short period orbit (generally combined with a small eccentric orbit due to the tidal effects of the host star).

The first rows of Table. 1.1 shows the RV semi-amplitude of the Solar System planets, evaluated from (1.2) in the perfect observational conditions ($i = 90^\circ$). For example, an Earth-like planet orbiting around a Sun at 1 AU (row 3) has a signal amplitude of 8.9 cm.s^{-1} whereas a Jupiter-like planet at 5.2 AU (row 5) produces a RV signal of 12.4 m.s^{-1} amplitude. To give an order of magnitude, a radial velocity variation of 12 m.s^{-1} corresponds to the detection of a wavelength shift in the spectral lines of $\Delta\lambda = 2 \times 10^{-4} \text{ \AA}$. Consequently, the detection needs an extreme stability of the spectrograph. Rows 9 and 10 illustrate K for an Earth-like planet in close orbit and a Super-Earth planet (at 1 AU) to be compared to Earth (row 3). When the planet is 10 times closer (or 5 times more massive), K is 3 (or respectively 5) times larger. The last three rows correspond to some known exoplanet signatures. First, 51 Pegasi b (row 11) evoked in Sec. 1.1 for

¹¹exoplanets.eu

¹² $0 \leq e \leq 1$, with $e = 0$ for a circular orbit and $e = 1$ for a parabolic orbit.

which the mass of the star is $M_\star = 1.04 M_\odot$ and the inclination of the orbit $i = 80^\circ$ leading to a RV semi-amplitude of 55.9 m.s^{-1} . As an illustration, the observed RV curve of 51 Peg b is shown in the left panel of Fig. 1.4 as function of the phase¹³. Follows α Centauri B b (row 12), which is a debated planet detection around a solar mass star ($M_\star = 0.907 M_\odot$) of spectral type K1V [Dumusque et al., 2012, Hatzes, 2013, Rajpaul et al., 2016], and Proxima Centauri b (row 13) orbiting a M dwarf star ($M_\star = 0.12 M_\odot$) of spectral type M5.5Ve [Anglada-Escudé et al., 2016].

Row number	Planets	Mean distance to the star (AU)	Periods (days)	Eccentricity	Mass (M_J)	K (m.s^{-1})
1	Mercury	0.38	88	0.2060	1.70×10^{-4}	0.008
2	Venus	0.72	225	0.0068	2.56×10^{-3}	0.086
3	Earth	1.00	365	0.0167	3.15×10^{-3}	0.089
4	Mars	1.52	686	0.0934	3.40×10^{-4}	0.008
5	Jupiter	5.20	4332 (11.9 yr)	0.0485	1.00	12.47
6	Saturn	9.58	10755 (29.5 yr)	0.0556	0.299	2.758
7	Uranus	19.20	30687 (84 yr)	0.0460	0.046	0.299
8	Neptune	30.05	60190 (164 yr)	0.0110	0.054	0.281
9	Earth in close orbit	0.1	36.5	0.0167	3.15×10^{-3}	0.28
10	Super-Earth	1	365	0.0167	15×10^{-3}	0.45
11	51 Peg b	0.052	4.2	0.013	0.46	55.94
12	α Cen Bb	0.04	3.23	0 (fixed)	3.55×10^{-3}	0.51
13	Proxima Cen b	0.048	11.18	< 0.35	4×10^{-3}	1.4

Table 1.1 – Orbital characteristics and RV semi-amplitudes of the Solar System planets evaluated for $i = 90^\circ$ (rows 1-8), a close-orbit Earth-mass planet (row 9), a Super-Earth of $5 M_\oplus$ at 1 AU (row 10), 51 Pegasi b (row 11), α Centauri B b (row 12) and Proxima Centauri b (row 13).

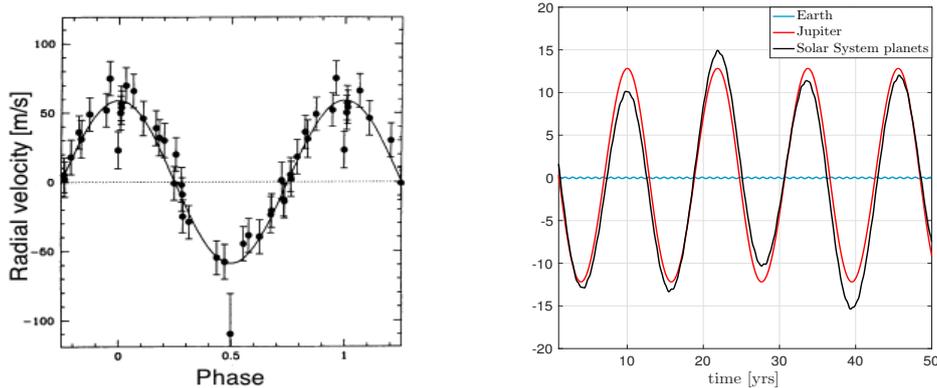


Figure 1.4 – Left: Observed radial velocity curve of the planet 51 Pegasi b (extracted from [Mayor and Queloz, 1995]). Right: RV signatures of the Earth (blue), Jupiter (red) and the entire Solar System planets (assuming no interaction between them, black) evaluated for $i = 90^\circ$. The RV of the entire Solar System is completely dominated by the Jupiter’s influence, which with Saturn, contains more than 90% of the whole Solar System planets mass.

¹³The phase is $\frac{t-t_0}{T_p}$ with t the time, t_0 the time of the planet passage at periastron, and T the planet period.

Considering now the orbit characterization, the signatures for N_p planets with Keplerian orbits can be expressed by:

$$V_R(t) = \sum_{p=1}^{N_p} K_p \left(\cos(\omega_p + a_p(t)) + e_p \cos(\omega_p) \right) + V_0, \quad (1.3)$$

with K_p the semi-amplitude given by (1.2), e_p the eccentricity, ω_p the argument of periastron¹⁴ ($0 \leq \omega_p < 2\pi$), V_0 the velocity of the system barycentre, and $a_p(t)$ the true anomaly. In practice, only the Keplerian parameters T_p , ω_p , e_p , M_p , and K_p can be derived from the observations. The shape of the RV planet signature in (1.3), with respect to the different Keplerian parameters, will be studied in details in Chap. 2 (Sec. 2.3).

In (1.3), the true anomaly is defined by Eq.(2.7), p.10 of [Perryman, 2011] as

$$a_p(t) = 2 \arctan \left[\sqrt{\frac{1+e_p}{1-e_p}} \tan \left(\frac{E_p(t)}{2} \right) \right], \quad (1.4)$$

with $E_p(t)$ the eccentric anomaly. The eccentric anomaly can be calculated using the mean anomaly $M(t)$ which is, at a given time, the angular distance of the planet from periastron:

$$M_{E,p}(t) = 2\pi \frac{t - t_0}{T_p} = E_p(t) - e_p \sin(E_p(t)), \quad (1.5)$$

with t_0 the time of the planet passage at periastron. The mean anomaly can be seen as a fictive mean motion around the planetary orbit and can be inferred directly from the data (see Fig. 1.5). The second equality in (1.5) is called the *Kepler equation* and has to be solved by iterative numerical methods such as the Newton-Raphson algorithm. At a given time, the planet position (indicated by the blue circle in Fig. 1.5) can be described both in terms of the true anomaly a_p (with respect to the ellipse centre, i.e. the system center of mass, here indicated by the star symbol) and the eccentric anomaly E_p (with respect to the auxiliary circle, i.e. the circle which circumscribes the planetary ellipse).

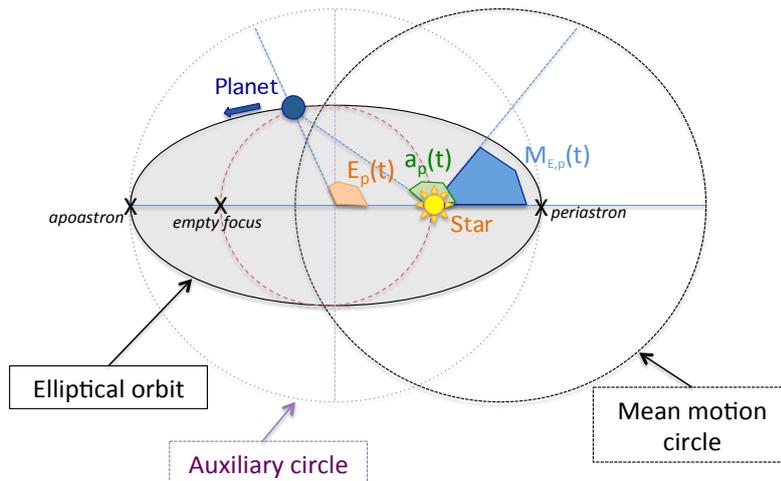


Figure 1.5 – Illustration of the three anomalies associated to an elliptical orbit.

¹⁴The periastron is the point where the planet is the closest to the star. The argument of periastron is the angle between the direction of the ascending node and the direction of the periastron.

As seen in Sec. 1.1, the detectability of the RV planet signature depends on the planet’s orbit inclination on the light-of-sight. The inclination parameter is involved in (1.3) through $K_p \propto M_p \sin(i)$ (cf. (1.2)) and cannot be determined from RV observations (but by transit for example). This leads only to a lower limit of the planet mass when the planet is only detected by the RV technique.

Fig. 1.6 shows the effect of the inclination on the RV curves for three different configurations:

- a) The “pole-on” geometry ($i = 0^\circ$), where the Doppler signature is minimum ($V_R = 0$). In this configuration the planet signature is not detectable and we have to use astrometric measurements to detect the planet.
- b) An intermediate case ($i = 45^\circ$), where the Doppler signature increases.
- c) The “edge-on” geometry ($i = 90^\circ$), where the RV semi amplitude is maximum ($\max V_R$). This is the best configuration.

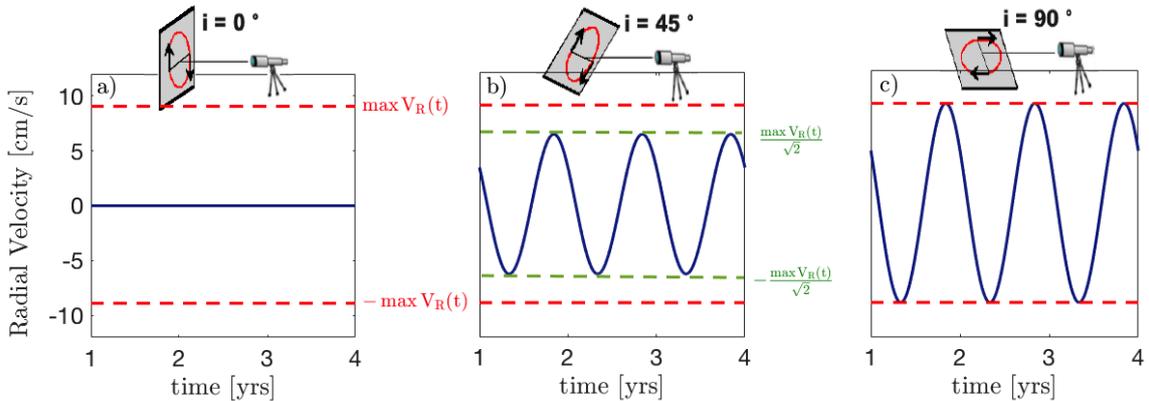


Figure 1.6 – Effect of the orbit orientation on an Earth-like planet RV signature (the Keplerian parameters are indicated in the row 3 of Table. 1.1).

1.2.2 Method of measurement and instrumental errors

As discussed in the beginning of this section, RV measurements are based on the systematic measure of Doppler shifts in the stellar absorption lines. Ignoring the relativistic effects ($V_R \ll c$), the velocities can be determined by the classical form of the Doppler effect given in (1.1). According to Hatzes & Cochran, the RV precision (or dispersion, σ_{instr}) is related to the resolution R of the spectrograph, the signal-to-noise ratio (SNR) S/N , and the spectral range B_λ of the spectrograph as [Hatzes and Cochran, 1992]:

$$\sigma_{instr} \propto R^{-1} \times (S/N)^{-1} \times B_\lambda^{-1/2}.$$

The resolution and the spectral range of the spectrograph fixed, the instrumental precision depends on the SNR. The SNR depends on a lot of parameters, which modify the saturation level (e.g. the signal magnitude, the detector characteristics, the stellar magnitude and spectral type, the exposure time of the observations, the considered wavelength range or the quality of the atmosphere). Achieving a RV precision of some m.s^{-1} is possible for two reasons.

First, a long-term instrumental stability, both in temperature, pressure and humidity, over months or years. This is possible with *échelle spectrographs*¹⁵ operating at high resolution.

¹⁵The principle of any spectrograph is to disperse the received starlight into its wavelength components. *Échelle spectrographs* use a reflecting grating (i.e. a dispersive element), composed of equally spaced grooves (facets)

Today (2017), the more precise spectrograph is HARPS, which is installed on the ESO 3.6-m telescope, La Silla, since 2003. HARPS operates in the optical domain (378 – 691 nm) with a resolution of $R = \frac{\lambda}{\Delta\lambda} \approx 120\,000$. According to the extreme stability and the characteristics of HARPS (e.g. captors and telescopes sizes, spectrograph resolution, calibration by Thorium-Argon lamps), and considering a photon noise limit evaluated in nominal condition (e.g. for stars with $M_V = 7$), this spectrograph is able to provide long term (1 month) wavelength reference allowing to achieve a precision around 1 m.s^{-1} until the saturation limit of the CCD is reached (300 000 electron peaks intensity) [Bouchy et al., 2001, Lovis et al., 2006]. This limiting value of 1 m.s^{-1} has to be seen as the intrinsic scatter on the observed RV. Signal with amplitudes below such limit are undetectable (whatever the number of observed data).

The detection also takes benefit from the instantaneous combination of many absorption lines in the stellar spectrum through the cross-correlation technique, which improves the measured SNR of the target Doppler shift (e.g. [Allende Prieto, 2007]). To do that, stars with spectral type latter than *F6* are the privileged targets as, at a given metallicity, they present a lot of absorption lines (see Fig. 1.7). For earlier-type stars, complications appear as they are often fast rotators making their (few) spectral lines wider (and consequently the wavelength shifts is less detectable). One notes however that adapted crossed correlation techniques using specific templates (e.g. [Galland et al., 2005]) or other dedicated techniques (e.g. [Astudillo-Defru, 2015]) exist for these stars

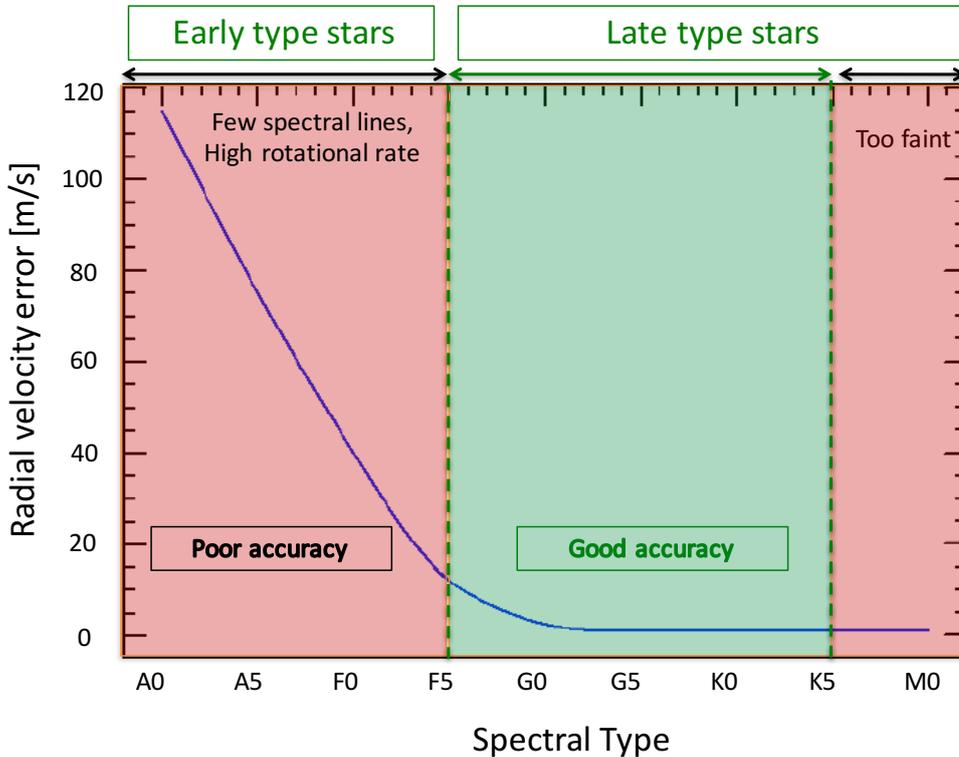


Figure 1.7 – RV accuracy depending on the spectral type. Early type stars, from A0 to F5, and late type stars, from K5 to M0, lead to poor RV accuracy. Main sequence stars, from F5 to K5, are more favoured for RV technique. Figure adapted from [Bozza et al., 2016].

leaned to a specific angle (the blaze angle) with respect to the grating normal. This angle of diffraction is chosen to make constructive interference and to diffract most of the light into high order (as a zero order corresponds to the white light which bring no information). One can report to [Vogt, 1987] for details on this instrument.

Following the Hatzes’ chapter in [Bozza et al., 2016], the more common methodology of RV extraction is based on the cross-correlation technique. The main steps are:

1. **Pixel measurements of the absorption lines position** in the stellar spectrum by an échelle spectrograph (e.g. HIRES, HARPS). In the spectrograph, many diffraction orders disperse the light and are recorded in a CCD where the absorption lines can be measured. The accuracy of the measurements depends on the instrumental resolution and on the target star. For example, Fig. 1.7 (extracted from [Bozza et al., 2016]) shows an estimate of RV errors as a function of the stellar spectral type (evaluated by taking into account the stellar mean rotational velocity and approximate stellar line densities). Cool stars (spectral type later than F6) have a good accuracy while more massive stars with high rotational rates and high effective temperature induce large RV errors.
2. **Conversion in wavelength shift** with a calibration method. The first technique consists in calibrating the observed spectrum with a simultaneous calibration done with a Thorium lamp. This is the method used with spectrographs like ELODIE and HARPS [Bouchy, 2015]. Another techniques consist in using an absorption cell [Griffin and Griffin, 1973] as for the spectrograph of the McDonald Observatory (US), or in using the Earth atmosphere as a reference (through telluric spectral lines). All these calibration methods need high stability to minimize the instrumental drifts. Some errors can happen during the data calibration step and one notes that a future promising method, the “Laser frequency Combs”, is in development for the next generation of spectrograph like ESPRESSO [Pepe et al., 2010].
3. **Cross-correlation of the entire spectrum.** Thousands of lines are measured in the spectra of FGK stars and the wavelength shift is deduced by cross-correlation of the entire spectrum with templates [Baranne et al., 1996]. Again, this is only possible by combining thousands of lines. Indeed, since HARPS’ camera pixel size is about $15 \mu\text{m}$ (corresponding to 800 m.s^{-1}), it would be impossible to measure a 1 m.s^{-1} displacement with a unique spectral line.
4. **Translation in radial velocities.** Using (1.1), the wavelength shift is measured with respect to (w.r.t.) the reference frame of the observer. As the referential of the measurement is linked to the spectrograph, we have to correct the shifts due to the Earth motion depending on the instrument location and epochs of measurements. These barycentric corrections include (among other things) the Earth’s orbital motion (up to 30 km.s^{-1}), the Earth’s rotation ($\approx 4.6 \text{ km.s}^{-1}$), the Solar System motion, and the proper motion of the star w.r.t. the Solar System (some milli arcseconds per year for the majority of the stars¹⁶ [Rouan, 2011]). These motions are evaluated with the ephemerides, the observer position, the distance to the Earth centre and the position of the observed star.

During the process of RV extraction, instrumental noises can hide the searched planet signatures. Fig. 1.8 shows the RV semi-amplitude K_p evaluated by (1.2) as function of the planet periods T_p (for $e_p = 0$ and $M_\star = 1 M_\odot$) for different planet masses ($1 M_J, 5 M_\oplus, 1 M_\oplus$). The two RV amplitudes indicated at right represent the limit range of the most accurate spectrograph in activity, HARPS (blue shade region), and the future spectrograph ESPRESSO (yellow shade region). One can observe that Earth-mass planets are not detectable with HARPS (at least for period > 1 hour), which is only able to detect Super-Earth with close orbit (period ≤ 1 year). On the contrary, the future high-precision spectrograph ESPRESSO will decrease the instrumental noise level by a factor of 10 allowing the detection of Earth-like planets (e.g. the

¹⁶Barnard’s star is the star known with the highest proper motion [Barnard, 1916]. This stars moves at $10.3''/\text{year}$ leading to observed RV variations around 110 km.s^{-1} . To keep an order of magnitude in mind, it would take around 175 years to this star to drift an angular distance equal to the diameter of the full Moon.

blue curve corresponding to $1 M_{\oplus}$ with 1 year period is at the expected detection limit of ESPRESSO). These performance will be achieved thanks to the reduction of the instrumental errors (by the spectrograph stability) and the improvement in the data calibration (through the laser frequency comb technique).

For these modern stable spectrographs, the instrumental noise will not prevent the detection of telluric RV planet signatures. Indeed, another source of noise appears when looking for small planets: the noise created by the star itself, which becomes the main limiting factor at such low instrumental levels.

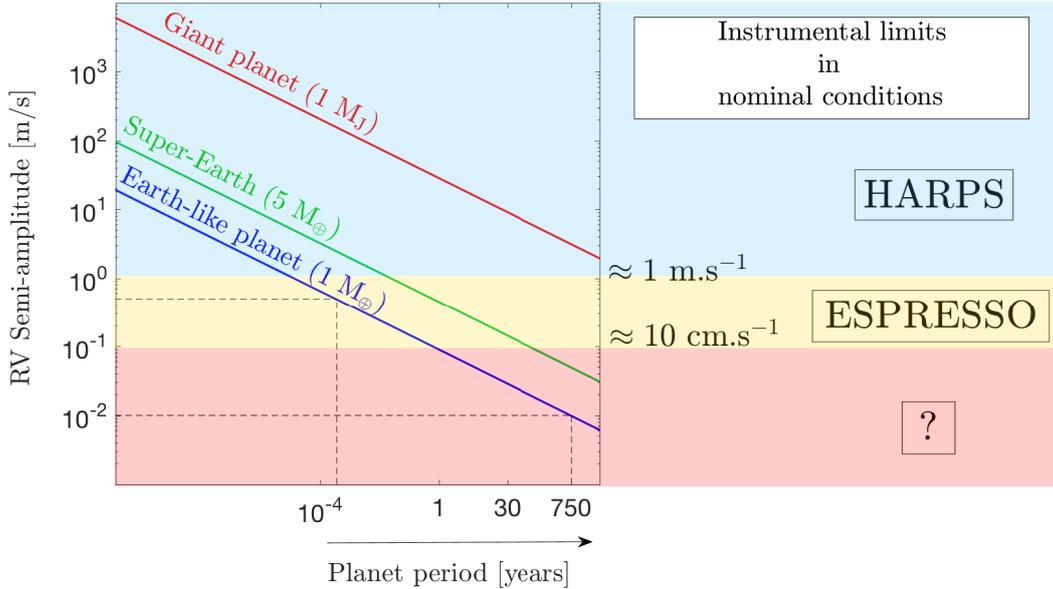


Figure 1.8 – RV planet semi-amplitude as function of periods for $e_p = 0$ and a solar mass star. The blue and yellow shade regions represent approximately the HARPS and ESPRESSO instrumental limits.

1.3 Stellar noises

Astrophysical noise sources arise from the stellar photospheric activity of the host star. Indeed, the light observed to detect the RV planet signatures is formed at the stellar surface and therefore is very affected by its dynamical environment. These plasma motions alter the shape of the stellar absorption lines and affect the RV under the form of correlated noises.

Ignoring these noises, or misunderstanding them, had lead to several false planet detections in the past (examples of debated detections are [Queloz et al., 2001, Figueira et al., 2010, Dumusque et al., 2012]). Among the stellar activity, one can distinguish three main sources of noise:

- the rotationally modulated starspots (dark areas) and plages (bright areas),
- the stellar oscillations (periodic light variations caused by resonant modes),
- the stochastic convection (granules).

Of course, this is a non-comprehensive list and a multitude of other phenomena participate to the spectral lines alterations. For example, one can also note the motions induced by stellar

flares¹⁷ (see [Reiners, 2009], which study the case of an M dwarf star) but these processes are not as important as the three above.

These three main noise sources can well be seen in RV data of the best known star: the Sun. The left panel of Fig. 1.9 shows one year of RV solar data as a function of time, measured by the GOLF spectrograph on board the SoHO spacecraft. This NASA mission is unique as it has collected more than 20 years of (quasi) continuous observations. Contrary to ground based spectrographs, GOLF extracts the RV by measuring intensity contrast in the sodium Doublet lines [Garcia et al., 2005]. One can see RV magnitudes of several m.s^{-1} dominated by the three main sources of the solar activity. The right panel illustrates the estimated power spectral density (PSD) evaluated by the periodogram (i.e. the squared modulus of the Fourier transform of the time series, see equation (2.24) in Chap. 2). At high frequency ($\nu > 10^{-2}$ Hz), the power is approximately flat due to the instrumental noise: one can consider that in such PSD region, the noise is White Gaussian (purple region). At lower frequencies ($10^{-2} > \nu > 10^{-3}$ Hz), we can observe the bump due to the oscillation modes, then the convective pattern ($10^{-3} > \nu > 10^{-5}$ Hz) and the magnetic activity ($\nu < 10^{-5}$ Hz). The periods larger than 70 days have been filtered out in this data set (i.e. $\nu < 1.65 \times 10^{-7}$ Hz, yellow shade region). This solar velocity power, where no planetary signature is present (the periods of Mercure and Venus are superior to 70 days), clearly shows a strong frequency dependence. The stellar activity affects the whole frequency range and may disturb the planet detection. Details on the physical properties of these stellar activity sources and their implications for extrasolar planet detection are given below.

1.3.1 Magnetic activity

Physical description. The magnetic activity of stars produces starspots (dark areas) and plages (bright areas) at the stellar surface. They appear as magnetic flux emergence, migrate in latitude during a given stellar cycle, and may move across the disk as the star rotates. In such region, the magnetic field is 1000 times larger than the global stellar field. Spots are regions with low temperature in the photosphere (around 4500 K on the solar spots umbra and 5500 K in the penumbra region [Solanki, 2003]). The number of these surface inhomogeneities changes with the stellar cycle and can be very large during the cycle maxima (see the Sun pictures in Fig. 1.10 taken during two phases of the solar cycle). In practice, these magnetic spots appear, evolve in size, and can migrate both in longitude and latitude at the stellar surface before they fade away. They rotate with the star and change the shape of the observed line profiles, as illustrated in Fig. 1.11 (extracted from [Haywood, 2015]). The top row of Fig. 1.11 illustrates different spot locations on the stellar surface, while the bottom row shows the associated spectral line profile distortion as a function of the spot location. If a spot is present on the observed stellar disk, it changes the shape of the observed line profiles as it modifies the global balance between upflows and downflows. It induces an inhibition of the convective blueshift effect (i.e. there is a deficient emission at a given wavelength) [Dravins et al., 1981, Thomas and Weiss, 1992]. This inhibition is due to the suppression of the convection by the strong magnetic fields inherent to the spot. Consequently, regions with strong magnetic activity appear redshifted with respect to the quiet photosphere [Meunier et al., 2010, Plavchan et al., 2015, Meunier et al., 2017b].

Influence on the detection of exoplanets. The active regions are problematic for exoplanet detection as they persist for timescales of several stellar rotation periods (days to months). The RV contribution of these features can thus mimic the periodic signatures of interest [Queloz et al., 2001]. The amplitudes change with the stellar rotational velocity and the spot filling factor (i.e. the fraction of the visible stellar disk covered by spots). For FGK

¹⁷The stellar flares are a brief variation of the stellar brightness due to a sudden energy release. They often (but not always) go with coronal mass ejection.

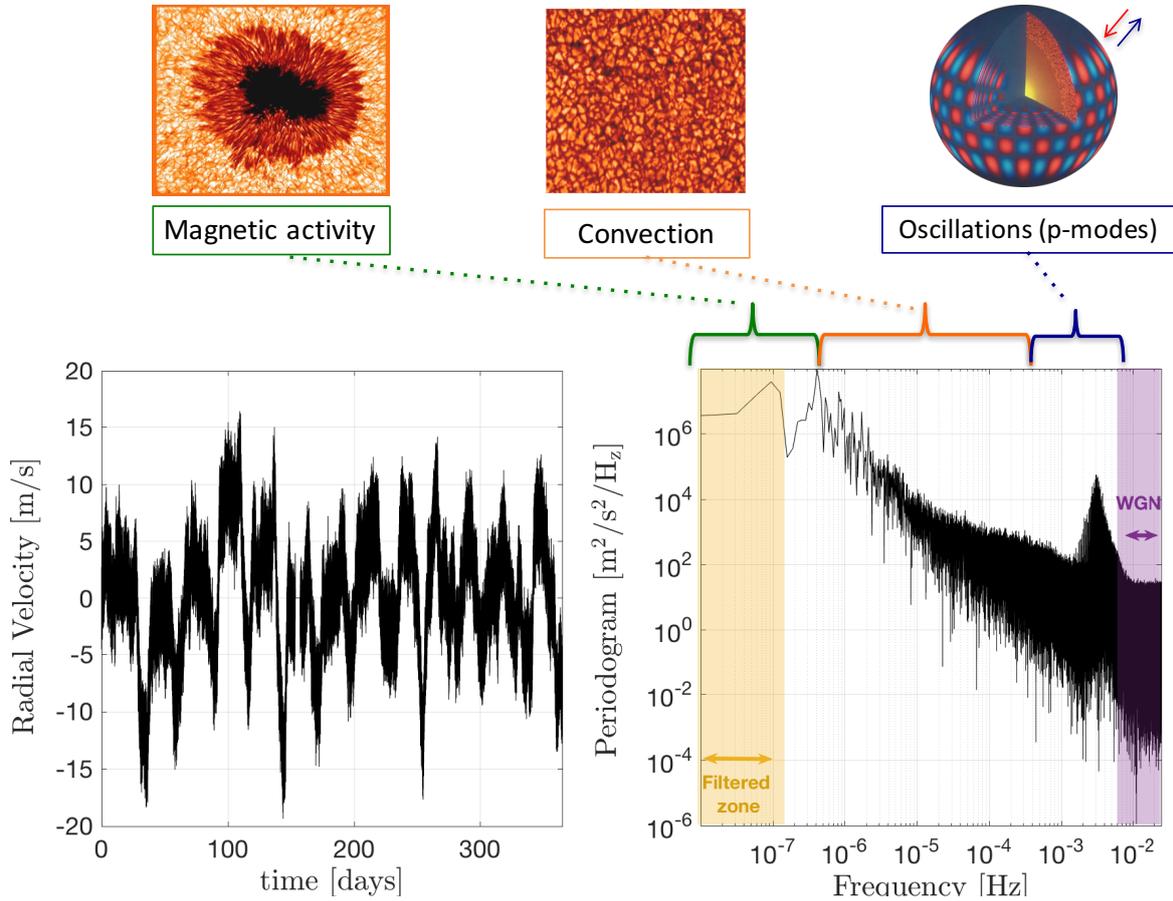


Figure 1.9 – Illustration of the solar colored noise. Left: Low frequency filtered solar RV times series taken by the GOLF spectrophotometer on board the SoHO satellite (D. Salabert, private communication). This time series of one year is (quasi) regularly sampled with a time sampling step of $\Delta t = 20$ seconds and has been taken during a solar minimum (1996). Right: Associated PSD estimate (evaluated through the periodogram Eq.(2.24)). From high to low frequencies, one can observe the different dominant components: the instrumental noise (purple shade region), the oscillation modes, the convection and the magnetic activity. This set of data has been filtered out at 70 days and frequencies $\nu < 1.65 \times 10^{-7}$ Hz should not be considered (yellow shade region). This figure shows the large RV variations and the clear frequency dependence of the solar activity.

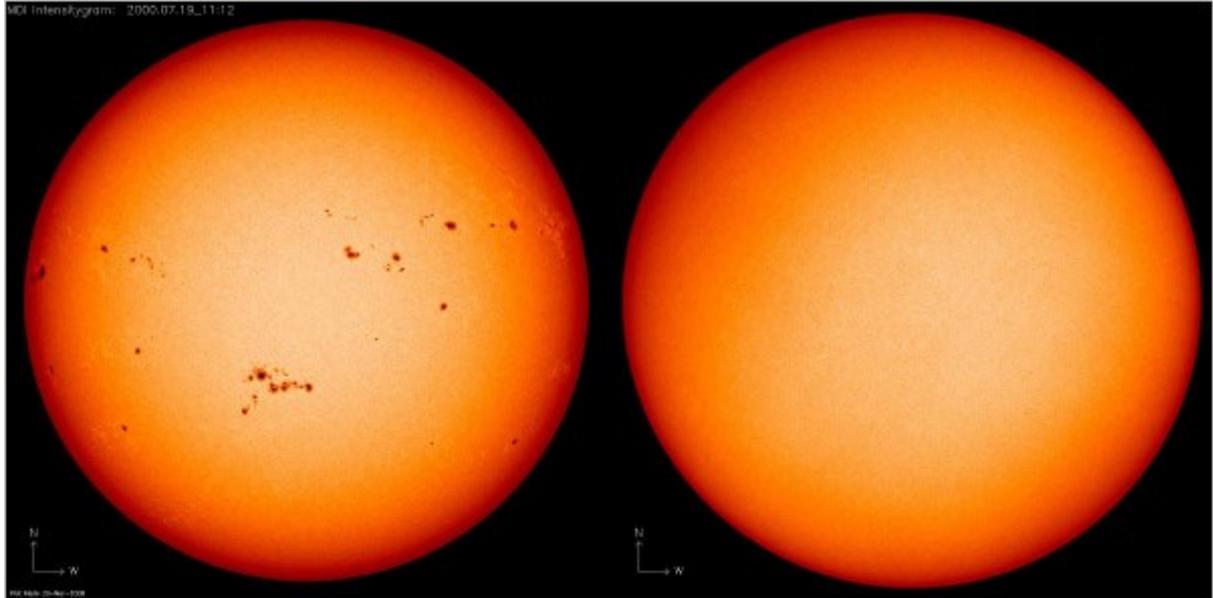


Figure 1.10 – Image of the solar disk in visible light taken by the instrument MDI onboard the SoHO spacecraft. At left, the image is taken during a solar maximum (on July 19, 2000) and at right, on solar minimum (March 18, 2009). At maximum, one can observe a multitude of stellar spots at the surface, in opposition to the quiet photosphere observed during the solar minimum. Credit: NASA.

stars, the typical spot effects are not superior to 2 m.s^{-1} , but it is enough to mimic or hide the smallest planetary signatures [Meunier et al., 2010]. Studying the Sun, Meunier et al. have determined empirical relationships between stellar parameters and the level of spot-induced RV jitter. These unresolved features create a very complex radial velocity signature that requires the use of activity diagnostics. For example, some common diagnostics used to infer the presence of the spot and plage signatures in RV data are:

- Measurements at different wavelengths to investigate the temperature contrast in spots and plages. For example, observations in the infrared instead of in optical, should reduce the amplitude of the spots signatures as the contrast between the cool spots and the hot photosphere decreases with increasing wavelengths, while the RV amplitude due to a planet should stay constant [Ballerini et al., 2012]. This trick was for example illustrated on solar observations used as a proxy [Marchwinski et al., 2015] but also to detect planets (e.g. a $10 M_J$ planet at 3.56 days was detected by combining near-infrared and millimeter-wavelength measurements [Setiawan et al., 2008]).
- Spectral indices measurements, like the Ca II and H&K lines, present an emission feature in the line's core due to chromospheric magnetic activity [Baliunas et al., 1995, Boisse et al., 2011, Meunier and Lagrange, 2013]. For example, this technique was used to invalidate the planet detection of HD 99492 [Kane et al., 2016]. For M dwarfs, the Balmer $H\alpha$ can also be used, as it presents correlations between the stellar activity and the equivalent width of the $H\alpha$ line. As another example, the detection of GJ 581 was debated as clear correlations between the $H\alpha$ line variations and the optical observations were found at the claimed planet period of 66 days [Robertson et al., 2014].
- Bisectors, measuring the spectral lines shapes, which are distorted by the presence of spots (but also by the oscillations and convection pattern). The resulting shape between two arbitrary chosen points in the resulting bisector can be analyzed to deduce the presence of spots [Hatzes and Cochran, 2000, Desort et al., 2007, Boisse et al., 2011].

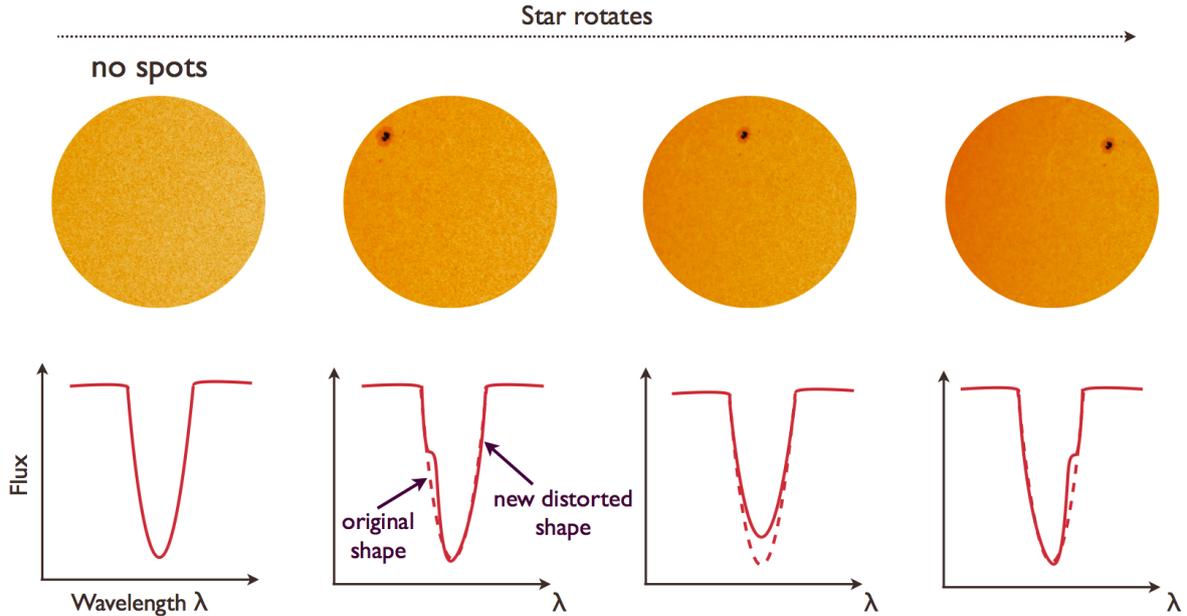


Figure 1.11 – Illustration of the distortion of the spectral line shape due to the presence of a starspot moving at the stellar surface. At the spot location, some light is removed which affects the centre and shape of the observed spectral line with period related to the stellar rotational velocity. This line asymmetry produces periodic variation in RV, which can mimic planetary signatures. Figure extracted from [Haywood, 2015] (p.15).

- Sine fitting at harmonics of the stellar rotational period to avoid the persisting rotationally modulated starspots and plages signatures [Dumusque et al., 2012].
- Simultaneous photometry and spectroscopy measurements to estimate the stellar perturbations. This method was introduced by Aigrain et al., and is known under the name of FF' method [Aigrain et al., 2012].
- Model with dark rotationally modulated spots (e.g. SOAP numerical simulations [Boisse et al., 2011]).

1.3.2 Oscillations

Physical description. Stars with an outer convective envelope are able to excite acoustic modes, also called pressure-modes or p-modes, at the surface [Stein, 1967, Stein and Nordlund, 2001]. Their properties are determined by the speed of sound profile inside the star interior, which makes them very interesting probes of the stellar interiors. They were observed for the first time on the Sun in 1960, with a period of 5 minutes and amplitudes of $\approx 20 \text{ cm.s}^{-1}$ [Leighton, 1960]. A few years later, this gave birth to heliosismology [Deubner and Gough, 1984]. Similar oscillations were detected in 2001 in (nearby) stars, other than the Sun (see the detection of oscillation modes on $\alpha\text{Cen A}$ by the spectrograph CORALIE [Bouchy and Carrier, 2001]). Now, we detect them in thousands of stars thanks to CoRoT and Kepler. For Sun-like stars, these oscillations have typical periods of few minutes and they produce a characteristic comb like structure in the stellar noise PSD (see the bump around 3 mHz in Fig. 1.9) [Bedding, 2014]. In radial velocity, the individual amplitudes of these modes are small (around 20 cm.s^{-1}).

The typical periods vary across the Hertzsprung-Russell diagram with the fundamental stellar

parameters following the relation

$$T_0 \propto \frac{\sqrt{T_{eff}}}{g}$$

with T_0 the oscillations period, T_{eff} the effective temperature and g the surface gravity of the star [Belkacem et al., 2011]. In general, the RV amplitudes and periods of the stellar p-modes increase with the temperature for dwarf stars [Dumusque et al., 2011]. For FGK stars, the modes lead to RV variations of magnitude ranging from some centimetres to meters per second depending on the stellar parameters [Lovis et al., 2006]. The period of the p-modes can reach some hours in Red Giants and the generated RV amplitudes can be 10 to 100 times larger than for Sun-like stars [Christensen-Dalsgaard, 2014]. One can report to [Bedding, 2011, Brown and Gilliland, 1994, García, 2015] for detailed reviews on the physical description of the stellar oscillation modes.

Influence on the detection of exoplanets. For main sequence stars, the oscillations do not represent a significant source of noise for extrasolar planet detection because their signatures are concentrated around periods of some minutes. In practice, attempts to “cancel” this noise consist in averaging the signal during exposure times longer than the oscillation periods. Typically, an exposure time of 10 – 35 minutes is used for Sun-like stars [Dumusque et al., 2011]. For giant stars however, they can become a more important source of noise for planet detection.

1.3.3 Convection

Physical description. The solar convection was observed for the first time on the Sun by Herschel in 1801 [Herschel, 1801]. Herschel observed at the solar surface hot fluid cells of gas moving due to buoyancy that were called later on *granules* [Dawes, 1864]. These granules are the manifestation of surface temperature inhomogeneities, the granulation. This surface granulation pattern is associated to heat transport by the convection on horizontal scales of several thousands kilometres [Nordlund et al., 2009]. The hot upflows carry the excess of entropy from below, reach the surface, and then cool by radiation and sink down into the intergranular lanes [Stein and Nordlund, 1998] (see Fig. 1.12). These motions influence the lines in many aspects: shifts, width, and depth. It creates also an asymmetry in the line since the blue wing is sensitive to upflows (granules) and the red wing to downflows (intergranular lanes) which are very asymmetric (e.g. [Asplund et al., 2000, Asplund et al., 2004, Asplund et al., 2009]). The granules cover the majority of the photosphere area and cause a global line shift towards blue, called *convective blueshift* for the lines that are formed at the photosphere. This asymmetry is correlated to the granules velocities and temperature (e.g. [Dravins et al., 1981, Dravins, 1987, Gray, 2005]) and is generally measured using line bisectors (e.g. [Dravins and Lind, 1984]). A granule alone has vertical motions of several hundreds m.s^{-1} . However, when looking to the entire stellar disk, which is composed of million of granules, the global RV variations of the granules and intergranular regions tend to compensate each other and lead to fluctuating velocity amplitudes of some meters (the velocity decreases approximately with the square root of the number of granules [Kjeldsen and Bedding, 2011]). This is, for example, observed by HARPS-N [Dumusque et al., 2015] or GOLF [Garcia et al., 2005] (see left panel of Fig. 1.9).

In the Sun, convective noise was first studied by Harvey, who identified different components as a function of their scales and periods, namely (from high to low frequencies) [Harvey, 1985]:

- the granulation: the smaller pattern (periods of some min or hours),
- the mesogranulation: the intermediate scale (periods of some hours to days),
- the supergranulation: the largest scale (periods of hours-days).

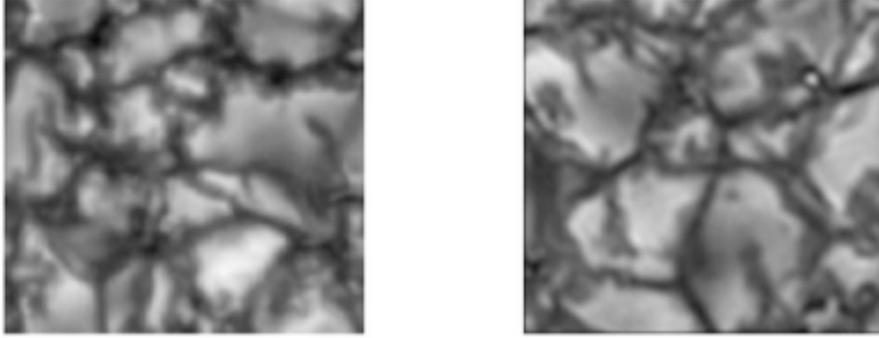


Figure 1.12 – Comparison between the emergent intensity at the solar surface coming from observations (right, Solar Swedish Telescope, Hizberger J., private communication) and simulations (left) [Bigot and Thévenin, 2008]. Dimension: 6000 km × 6000 km.

Harvey derived empirical laws to fit the stellar power spectrum part dominated by such process (cf. Fig. 1.9). The description of the different convective patterns are based on (Lorentzian) functions fits, which are parametrized by the different amplitudes and timescales of the involved phenomena. The convection parts in the Harvey’s functions are commonly described as a sum of Lorentzian:

$$F_H(\nu, \boldsymbol{\theta}_n) = \sum_{i=1}^n \frac{4\sigma_i^2 \tau_i}{1 + (2\pi\nu\tau_i)^{\beta_i}}, \quad (1.6)$$

with ν the frequency, $n \in [1, 3]$ the number of considered components, and $[\boldsymbol{\sigma}^2, \boldsymbol{\tau}, \boldsymbol{\beta}]^\top$ the convection parameters to fit (with σ^2 the rms velocities, τ the characteristic timescale, and β the decay rate of the considered component).

At the smallest scale, the granulation has an averaged lifetime of about 10 – 15 min for the Sun and each granule has a characteristic size around 1000 kms [Title et al., 1989, Nordlund et al., 2009]. The existence of a mesogranulation convection scale, initially reported by November et al., is a source of debate [November et al., 1981, Chou et al., 1991, Rieutord et al., 2000], and in practice the “Harvey function fit” appears to work well without this intermediate component (e.g. [Lefebvre et al., 2008] succeeded in fitting the GOLF solar PSD with only two components). At the largest scale, the supergranulation has size around 30 000 km and a typical life time of ≈ 17 hours [Del Moro, 2004, Williams and Pesnell, 2011, Williams and Pesnell, 2014]. The properties of the solar granulation and supergranulation are listed in Table. 1.2 below. Fig. 1.13 shows three solar Dopplergrams¹⁸ (i.e. velocity map) taken by *Michelson Doppler Imager* (MDI) on board SoHO during a solar minimum (March, 1996). The light and dark regions represent respectively the up and down motions of the gas. In the top panel, one can see the solar rotation with velocity around 2 km.s⁻¹, which appears as a shift from dark to light (with the dark color indicated a motion toward the camera). In the bottom panel, the rotation of the Sun has been subtracted by a polynomial fit and an average of 30 images taken with a sampling rate of 45 seconds has been made to illustrate the solar supergranulation. One can observe velocities of the convective pattern with maximum values around 400 m.s⁻¹. The observed velocity is lower in the disk centre as the velocity is perpendicular to the light-of-sight of MDI. The last panel shows the granulation (1 image obtained by removing the supergranulation and rotation components). One can observe the small convective pattern at the surface with velocities around 600 m.s⁻¹ (this value is slightly underestimated due to the spatial resolution).

¹⁸Source: <https://sohowww.nascom.nasa.gov/gallery/HeliOSEISMOLOGY/>

	Granulation	Supergranulation
Average size	$\approx 1\,000$ km	$\approx 30\,000$ km
Average speed	≈ 600 m/s	$\approx 100 - 150$ m/s
Average lifetime	10 – 15 min	≈ 17 hours

Table 1.2 – Properties of the solar granulation and supergranulation.

The convection is present at the surface of all late type stars (FGK stars, see [Magic et al., 2013, Trampedach et al., 2013]). The characteristics of convection timescales and velocity amplitudes change with the fundamental stellar parameters (T_{eff} , $\log g$, $[\text{Fe}/\text{H}]$). The granulation velocities increase with the temperature for dwarf stars, i.e. they decrease from G subgiants to K dwarf stars [Dravins, 1990, Dumusque et al., 2011, Meunier et al., 2017a]. The granulation at the surface can not be observed by resolved observations as for the Sun and astronomers have to rely on the spectral lines in terms of width, shapes and strength variations to characterize it.

Based initially on the Sun study, reliable hydrodynamical simulations have been developed, which brought a lot of informations on the granulation activity (see Sec. 1.4). If the signal-to-noise ratio of the spectral line is sufficiently large, the line asymmetries due to granulation are reproducible by using 3D magnetohydrodynamic simulations of the solar convection (see [Dravins, 1999, Asplund et al., 2000, Dravins, 2008] and [Cegla et al., 2013] for the context of extrasolar planet detection). For more details about convection, one can refer to [Nordlund et al., 2009, Rieutord and Rincon, 2010, Williams and Pesnell, 2011].

Influence on the detection of exoplanets. The granulation component signatures is a stochastic effect (through spectral line asymmetries), which is correlated over periods of some minutes to several days. In RV, the convection produces variations from some centimeters to meters by second [Dumusque et al., 2011]. Consequently, the convection reaches RV variations of the same order of magnitude as the telluric planets [Meunier et al., 2015]. It is clear that accounting for this stochastic noise is a fundamental issue to be able to reliably claim a planet detection at the level of a few tenth of cm.s^{-1} precision.

To date, few strategies exist to counteract the effects of this intrinsic noise in the context of exoplanet detection. Dumusque et al. propose an observational strategy to average out a part of the granulation contribution [Dumusque et al., 2011]. Through a decrease of the number of measurements, the authors showed that separating the number of observations by some hours and using a specific exposure time allows to average this noise. This method was successful in some cases (e.g. [Pepe et al., 2011]) and is the common methodology applied in practice.

However, reported to simulation studies, Meunier et al. shows that, even after several hours of smoothing, the convective jitter amplitudes and correlations remain significant and cannot be averaged out to neglect its contribution when looking for exo-Earth planets signals [Meunier et al., 2015]. The authors studied the effect of both the granulation and supergranulation components by generating RV simulations of a collection of cells with different granule sizes and lifetimes. The granules were generated using an empirical size distribution for which they assigned a lifetime, a position on the disk, and a RV and intensity contributions. They found that both granulation and supergranulation are not completely averaged out by using the observation strategies proposed by Dumusque et al., and that further techniques have to be derived to deal with this noise source when looking for telluric planet signatures. Meunier et al. concludes that the convective noise sources has a significant impact on the

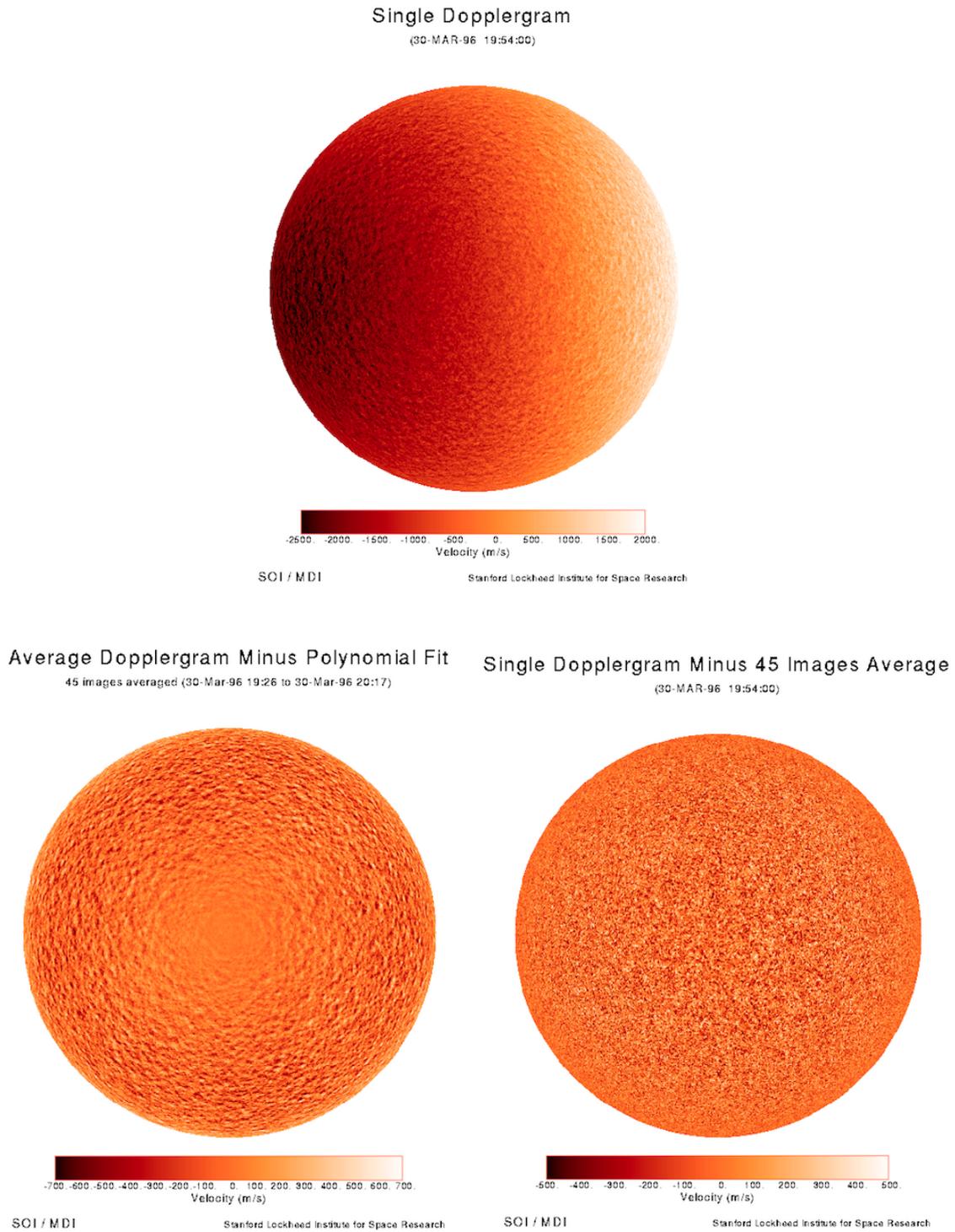


Figure 1.13 – Dopplergrams of the Sun taken by the MDI instrument on board SoHo. The top panel illustrates the row Dopplergram where the rotation appears as a shift from dark to light. The bottom panels show the supergranulation component (rotation removed, left) and the granulation (rotation and supergranulation removed, right). Credit: NASA.

detection limits even for Earth-mass planet signatures in the HZ of Sun like stars. This noise source has to be taken into account as well as the noise coming from the magnetically driven phenomena (as starspots and plages), which is the only noise source considered in the majority of the publications related to exoplanet detection.

One finally notes an interesting attempt to deal with the granulation component, based on the fact that recent 3D hydrodynamic simulations can reliably reproduce the spectral line asymmetries (cf. Sec. 1.4): Cegla et al. aims to localize specific lines significantly affected by the convection to exclude them in the RV extraction process [Cegla et al., 2013, Cegla et al., 2015].

1.3.4 Summary

To summarize, when looking for long period planets, the most annoying source of stellar noise limiting the planets detection are due to the magnetic activity (cool spots, hot faculae, and plage) as they produce RV variations at long periods (from days to years). Sometimes their effect is manageable by looking to stellar chromospheric indicators, sometimes not.

Oscillation modes induce a noise source, which is manageable by averaging the observations over several minutes for Sun-like stars. When looking for short period planets, the most significant noise source comes from correlations due to the stellar convection. This convective “noise” could produce false planetary detections and has to be taken into account. Moreover, numerical studies showed that the convective noise contribution could also be a limiting factor to detect exo-Earth.

Recently, a challenge under the form of a “blind test” was proposed to the community of planet hunters (cf. [Dumusque, 2016, Dumusque et al., 2017]). This test aimed to compare different strategies to detect planets in the instrumental and stellar colored noises. By injecting different synthetic planetary signals in simulated RV data at the 1 m.s^{-1} level (including instrumental and stellar noises), Dumusque challenged the community to find these signatures. The results show that the most efficient methods to find planets was to take into account the different activity indicators (e.g. Calcium and H&K lines) to identify the stellar activity and use red noise¹⁹ models to take into account the stellar noise correlations. However, these methods were not able to find the smallest planetary signatures and found a significant number of false positive. This proves that there is still the need of further investigations to deal with the stellar correlated noise. Understanding the stellar activity is a key point and one of the main challenge we have to face to be able to reliably claim the smallest planet detections.

1.4 Simulations of the convective stellar activity

In the 80’s, the first realistic models of the stellar convection activity were developed in 3 dimensions [Nordlund, 1982]. Such models are *ab initio* in the sense that they only need the information on the fundamental stellar parameters to solve equations of physics. This kind of hydrodynamic 3D code resolves equations of compressible radiative hydrodynamic in 3D, including radiative transfert and opacities (depending on the wavelength) and the Equation-of-states taking into account molecular ionisation, formation and dissociation. These 3D magneto-hydrodynamic simulations of stellar surface use cartesian box at the surface whose sizes (surface and depth) depending on the type of stars. They allow to simulate surface effects as the convection. A blueprint of a such local 3D box of simulation is illustrated in the left panel of Fig. 1.14.

Several codes exist, differing principally by the numerical scheme, e.g. the STAGGER code [Nordlund, 1982, Stein and Nordlund, 1998, Ludwig et al., 1999, Asplund et al., 1999,

¹⁹Power law noise

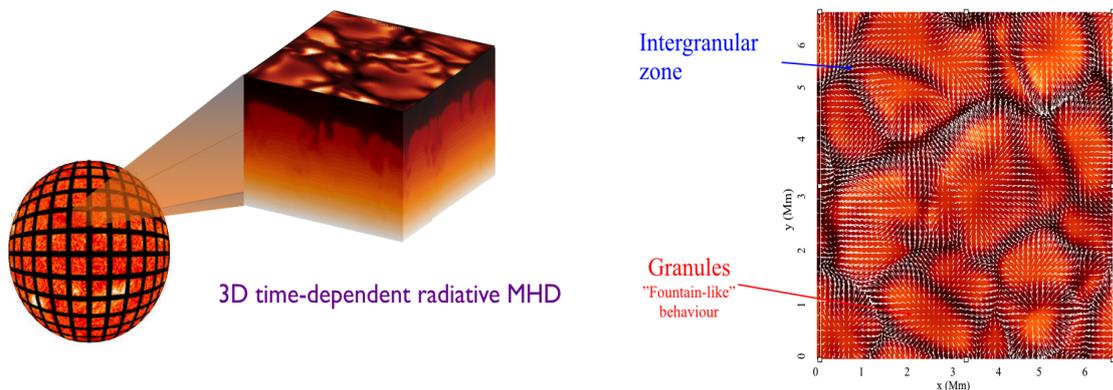


Figure 1.14 – Left: blueprint of a local box simulation (courtesy of L. Bigot). Right: Snapshot of one local box with the velocity field represented. One can observe, as expected, velocities in the radial direction in the centre of the granules and in the opposite direction in the intergranular zones.

Bigot et al., 2011], MuRAM [Vögler et al., 2005, Cegla et al., 2015] or CO⁵BOLD 3D [Freytag et al., 2012]. These were firstly used for stellar diagnostics. They showed a lot of corrections in relation to the hydrostatic 1D models. The main advantage is due to their hydrodynamical nature. They do not need adjustable parameters as micro or macro turbulence. In 1D, these fitting parameters follow an empirical law, which lead to model dependent results and strong uncertainties about chemical abundances. Indeed, the convection modifies the line shapes and consequently the chemical abundance (e.g. [Asplund et al., 2000, Asplund et al., 2004, Asplund et al., 2009]). They can produce synthetic spectral lines to generate the granules intensities and velocities (see the right panel in Fig. 1.14). Contrary to hydrostatic models in 1D, the 3D simulations can reproduce the surface inhomogeneities and so characterise the stellar convective noise.

For example, Fig. 1.12 shows a comparison between the emergent intensity at the solar surface coming from solar observations (right) and from such 3D simulations (left) [Bigot and Thévenin, 2008]. The realism of such simulation is evident. A result in velocity is shown in Fig. 1.15, which illustrates periodograms of simulated RV solar time series of 66 simulated days. These series have been divided into three different epochs to build 3 different periodograms. At low periods, one can observe the oscillations, which are naturally generated from the simulation. At the longer periods (low frequencies), one can observe the stochastic behaviour of the granulation whose manifestation in the periodogram is the apparition of random peaks at different frequencies. These variations can be wrongly interpreted as planetary signatures (see the peaks amplitudes variations at the frequency indicated by the dotted red lines). This figure clearly illustrates the need of the statistical characterization and control of this noise, as it can mimic or hide periodic signatures.

These simulations can be done for all late type stars (cf. [Dravins, 1990, Ludwig et al., 2002, Magic et al., 2013, Trampedach et al., 2013]). For example, Fig. 1.16 illustrates the emergent luminosity at the disk centre for different spectral type stars resulting from the STAGGER code. The granules sizes depend on the star through its temperature and gravity. The velocity of the convective cells increases with the temperature as the star becomes more opaque and the convection has to increase to evacuate the entropy excess to the surface. According to these simulations, Svensson & Ludwig show that the spectral power density (P) is strongly dependent on the stellar surface gravity (g) through [Svensson and Ludwig, 2005]

$$\log(P) \propto -\frac{1}{2}\log(g).$$

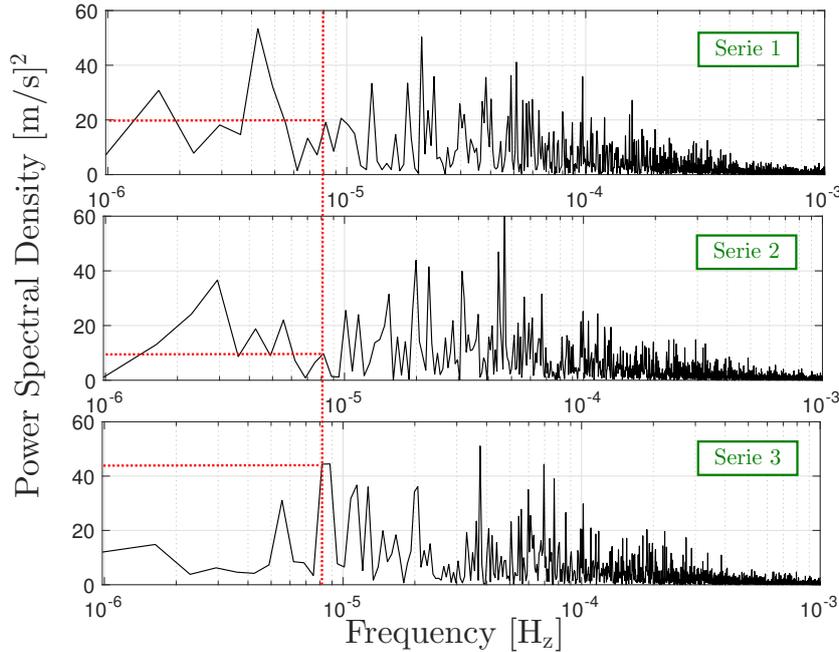


Figure 1.15 – Power spectral density estimate of simulated solar RV at three different epochs. The dot red lines indicate the peaks amplitudes at a frequency corresponding to 2 days. One can observe the stochastic behavior due to the convection, which can reach amplitudes of telluric planet RV signatures. Credits: Lionel Bigot.

1.5 Conclusions on the astrophysical context of the thesis

In the last two decades, the new field of exoplanet hunting has widely grown, with now more than several thousands of detections. While the majority of these detections are giants, Super-Earth or close-orbit planets, the quest of exo-Earth planets begins to be achievable (e.g. Gliese 667c [Anglada-Escudé et al., 2013]).

Among all the existing detection techniques, we focus in this thesis on the radial velocities. This technique is currently one of the most successful (with transits) and allows to determine the planet mass (up to a $\sin(i)$ factor). It is particularly efficient for cool low mass stars, short period and massive planets. In this method, an exo-Earth signature has an amplitude 1000 times lower than a Jupiter planet. These planetary signatures can be modelled thanks to the Kepler's equations and depend on 6 parameters: the planet mass, period, eccentricity, argument and time of passage at periastron and the barycentre velocity. By this technique, the detection of Earth analogue needs to face two major challenges:

1. The first is technological. In practice, the observed data result from a complicated velocity extraction process based on the observed stellar spectrum. Several sources of instrumental noise can perturb the measurements (e.g. temperature or pressure variations, calibration, guiding errors). Until the commissioning of the HARPS instrument, in 2003, the precision limit were due to the instrumental noise. This instrument is the most precise in activity today and allows to detect Super-Earth planets in close orbits. In the next years, the new generation of spectrographs such ESPRESSO will allow to decrease the detection threshold to some cm.s^{-1} making, in principle, small planets signatures detectable.
2. The second is astrophysical because at this level of instrumental precision, the surface activity of the host star creates a frequency dependent noise, which can hide or mimic the planet signatures. This stellar jitter affects the radial velocities in different ways:

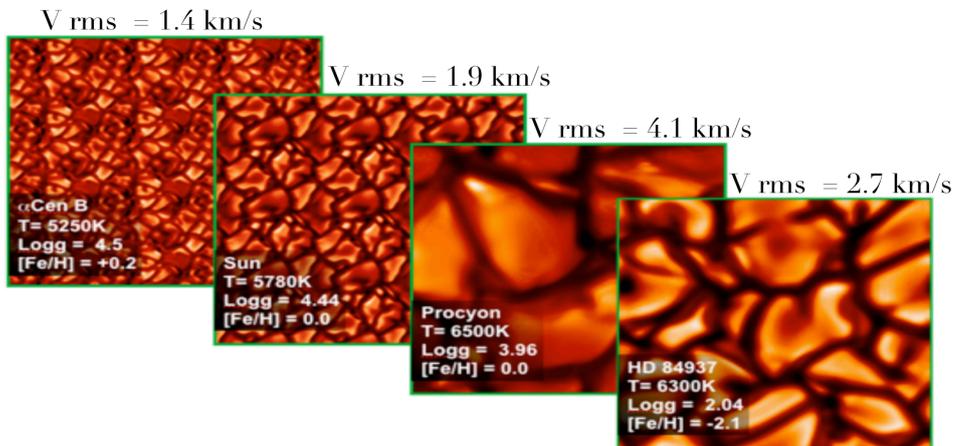


Figure 1.16 – Illustration of the granulation at the stellar surfaces. The different panels represent snapshots of different type of stars (up to the front): α Centauri B, the Sun, Procyon and HD 84937. In each case, the resolution grid is $128 \times 128 \times 96$ km corresponding to horizontal dimensions of 20000×20000 kilometers. The stellar parameters are shown in each panel. Courtesy: Lionel Bigot.

it has a stochastic part due to stellar surface convection and cyclic parts due to the magnetic activity (e.g. starspots) and oscillations modes. These noises have to be precisely understood to be able to claim real planet detections at the cm.s^{-1} levels.

This second point is the problem that motivated this thesis. While the magnetic part (e.g. spots, plages) can be potentially removed thanks to chromospheric indicators, the stochastic parts due to the convection process are not well constrained in practice. Stellar convection is a noise source always present at the stellar surfaces of all the late-type stars. This is an intrinsic physical limit to deal with to be able to detect the smallest planetary signatures at the sub-m.s^{-1} level. This thesis focuses on this colored stellar noise contribution.

To deal with this noise, the basis idea of my thesis is to couple a statistical approach with the knowledge of reliable 3D hydrodynamical (HD) simulations of such phenomenon in order to control the small planet detections. Simulations of the convective noise are now reliable and they can be useful tools to deal with the noise contribution due to the convection. The working hypothesis made in this work assumes that long series of such simulations can be done and used to design new and efficient statistical detection approaches. A first underlying advantage of using such HD simulations in the detection process is that the noise is modelled according to our knowledge of the physical processes (encapsulated in the HD simulations) and is not model dependent (as when using empirical functions).

Chapter 2

Detection of periodic signatures in noise: tools, models and tests

Contents

2.1 Hypothesis testing	30
2.1.1 Design of hypotheses	30
2.1.2 Statistical tests and decision rule	31
2.1.3 Frequentist approaches: tests based on likelihood functions	38
2.1.4 A brief comment on Bayesian approaches	40
2.2 Periodogram(s)	40
2.2.1 The classical periodogram	40
2.2.2 Periodogram of a single sinusoid	41
2.2.3 Parseval's identity	44
2.2.4 Periodogram resolution	45
2.2.5 Bias and variance of the periodogram	45
2.2.6 Asymptotic distribution in the case of a pure random process	45
2.2.7 Asymptotic distribution of P in the case of a periodic signal in noise	46
2.2.8 A note on some modified periodograms	51
2.3 Planetary signatures in the periodogram	52
2.3.1 Signatures in the time domain	52
2.3.2 Signatures in the Fourier domain	53
2.3.3 Discussion and signal model	57
2.4 Statistical tests applicable to the periodogram	58
2.4.1 Preliminary notations	58
2.4.2 Classical tests	59
2.4.3 Tests designed for multiple sinusoids	62
2.5 Adaptive tests that can be applied to the periodogram	66
2.5.1 Introduction of the problem	66
2.5.2 Example: Kolmogorov-Smirnov test	66
2.5.3 Higher Criticism	67
2.5.4 Berk-Jones test	70
2.6 Sinusoid detection in unknown noise statistics: related works	71
2.6.1 Non parametric approaches	72
2.6.2 Parametric methods	74
2.7 Conclusions	75
2.7.1 Main conclusions	75
2.7.2 This thesis	76

In this chapter, we present one basic methodology and tools for the problem of detecting periodic signatures in noisy data.

In the first part (Sec. 2.1), we explain the concept of *hypothesis testing* by introducing the basics steps to build a hypothesis test, how to measure the detection significance, and how to evaluate the test's performance. The considered hypothesis testing methodology will be focused on the frequentist approach but we will also mention the bayesian approach (yet without details, as it is not studied in this manuscript).

In the second part (Sec. 2.2), we introduce the most common tool to detect periodic signals, based on Fourier analysis: the periodogram. We will discuss the periodogram properties depending on the presence (or absence) of signal and depending on the noise statistics. We will also mention some alternative forms of the classical periodogram, introduced to counteract its drawbacks.

In the third part (Sec. 2.3), we study the properties of radial velocity signatures introduced in (1.3) in the Fourier domain. The results presented in this section have been published in [Sulis et al., 2016a].

In the fourth and fifth parts (Sec. 2.4-2.5), we present some statistical tests that can be applied to the periodogram depending on the knowledge (or not) of the searched signal parameters and the noise statistics.

We will conclude with the framework in line with my studies (Sec. 2.7), namely the detection of periodic Keplerian signatures in correlated noise.

Along these sections, we provide various illustrations of theoretical notions through examples. Those can be skipped for a reading focused on the main lines of the thesis.

2.1 Hypothesis testing

2.1.1 Design of hypotheses

In the middle of the 50's, Lehmann introduced the modern bases of hypothesis testing (see the recent review [Lehmann and Romano, 2005]). The hypothesis testing method consists in comparing between two (or more) possibilities for the data model. We speak about a *binary hypothesis testing problem* in the simple case of two hypotheses and about a *multiple hypothesis test* when there are more than two. In a “noise *vs* signal” framework, the goal of testing hypotheses is to use the available observations and the prior knowledge, if available, as efficiently as possible to make a decision between the presence or absence of a target signal within the data. Following [Kay, 1998], the construction of such tests can be divided into four steps:

1. State the hypotheses of the considered problem,
2. Identify the appropriate test statistic,
3. Specify its associated significance level by selecting a test threshold,
4. Make a decision.

In this manuscript, we mainly consider the case of a *binary hypotheses test*, characterized by two different statements:

- \mathcal{H}_0 : the null hypothesis, there is only noise within the data
- \mathcal{H}_1 : the alternative hypothesis, a signal is present within the noisy data.

As will be mentioned in Sec. 2.2 and 2.3, we will sometimes consider a multiple hypotheses framework, as in the Fourier domain multiple frequencies can be affected by exoplanetary signal(s).

2.1.2 Statistical tests and decision rule

For clarity, let us denote by X a random variable (r.v.) and by x its realizations. Vectors are denoted by $\mathbf{X} = [X_1, \dots, X_N]^\top$ and $\mathbf{x} = [x_1, \dots, x_N]^\top$.

Concept of statistical tests

To find evidence for or against \mathcal{H}_0 and be able to make a decision, one has to select an appropriate decision criterion through a detection test. A test statistic is a mathematical function of the data. Considering the data vector \mathbf{x} , the test statistic can be generally written as

$$T(\mathbf{x}) := T(x_1, \dots, x_N), \quad (2.1)$$

with $T(\cdot)$ a (deterministic) function.

In a hypothesis test, this test statistic is a decision variable. If \mathbf{X} is a random process, $T(\mathbf{X})$ is a random variable as well. For some data \mathbf{x} , $T(\mathbf{x})$ has to be compared to a decision criterion, $\gamma \in \mathbb{R}$, to make the test

$$T(\mathbf{x}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma. \quad (2.2)$$

The null hypothesis is then rejected if $T(\mathbf{x}) > \gamma$ and accepted if $T(\mathbf{x}) \leq \gamma$. Equation (2.2) is a *decision rule*. In the literature, this test is classified as a *one-tailed right-sided test* because $T(\mathbf{x})$ has to be superior to γ to decide \mathcal{H}_1 . When the test is $T(\mathbf{x}) \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\gtrless}} \gamma$, one speaks about a one-tailed left-sided test.

The threshold γ controls the significance level of the test, i.e. the probability of making wrong decisions or rejecting \mathcal{H}_0 when it is true. For example, a significance level of 0.01 indicates a 1% risk of error, i.e. of wrongly concluding the presence of a signal when there is only noise within the data.

To correctly determine the threshold's value, the statistical test should be characterized by a known probability distribution function (pdf) of $T(\mathbf{x})$ under \mathcal{H}_0 , noted $\varphi_T(t|\mathcal{H}_0)$. The pdf of $T(\mathbf{x})$ under \mathcal{H}_0 and \mathcal{H}_1 should be as different as possible for the test to be powerful. If the pdf is partially unknown under \mathcal{H}_0 , it is still possible to control the test's significance level under some conditions (as we will see in Sec. 2.4, Sec. 2.7, and Chap. 3).

To evaluate the performance of a particular decision rule, different statistical indicators exist. The main ones are the *detection/error probabilities*, the *P-values* and the *Receiver operation characteristics* (ROC) curves discussed below. The reader can report to Kay's book for more details about these indicators [Kay, 1998].

Error probabilities

The choice between the model hypotheses always leads to a probability of error and a probability of making the right decision. In the case of a binary hypotheses test, four probabilities can be defined for a test of the form (2.2), namely:

- The *probability of detection* (P_{DET})

$$\begin{aligned}
P_{DET} &:= \Pr(\text{“accept } \mathcal{H}_1\text{”} | \text{“}\mathcal{H}_1 \text{ is true”}) \\
&= \Pr(T(\mathbf{x}) > \gamma | \mathcal{H}_1), \\
&= \int_{\gamma}^{+\infty} \varphi_T(t | \mathcal{H}_1) dt,
\end{aligned} \tag{2.3}$$

- The *probability of false alarm* (P_{FA}), or *Type I errors*

$$\begin{aligned}
P_{FA} &:= \Pr(\text{“accept } \mathcal{H}_1\text{”} | \text{“}\mathcal{H}_0 \text{ is true”}) \\
&= \Pr(T(\mathbf{x}) > \gamma | \mathcal{H}_0), \\
&= \int_{\gamma}^{+\infty} \varphi_T(t | \mathcal{H}_0) dt,
\end{aligned} \tag{2.4}$$

- The *probability of missed detection* (P_M), or *Type II errors*

$$\begin{aligned}
P_M &:= \Pr(\text{“accept } \mathcal{H}_0\text{”} | \text{“}\mathcal{H}_1 \text{ is true”}) \\
&= \Pr(T(x) < \gamma | \mathcal{H}_1) \\
&= \int_{-\infty}^{\gamma} \varphi_T(t | \mathcal{H}_1) dt \\
&= 1 - P_{DET}.
\end{aligned} \tag{2.5}$$

The fourth case where \mathcal{H}_0 is retained whereas it is true, corresponds to $\Pr(T(x) < \gamma | \mathcal{H}_0) = 1 - P_{FA}$. These probabilities are listed in Table. 2.1 below.

	Claim \mathcal{H}_0	Reject \mathcal{H}_0
\mathcal{H}_0 true	Correct decision ($1 - P_{FA}$)	Wrong decision (P_{FA})
\mathcal{H}_1 true	Wrong decision (P_M)	Correct decision ($P_{DET} = 1 - P_M$)

Table 2.1 – Probabilities involved when evaluating a decision rule.

Before illustrating the distribution of a particular test, it is worth mentioning two useful distributions that will often appear in this manuscript: the central and non-central F distributions. Their properties are also listed in Appendix. A with the Gaussian and chi-squared (χ^2) distributions.

Definition 3. F-distribution

A F -distribution (defined also as *Fisher-Snedecor law*) corresponds to the scaled ratio of two independent chi-squared distributed random variables $V_1 \sim \chi_{d_1}^2$ and $V_2 \sim \chi_{d_2}^2$ with d_1 and d_2 the degrees of freedom [Abramowitz et al., 1972]

$$\frac{V_1/d_1}{V_2/d_2} \sim F(d_1, d_2).$$

The probability density function of a r.v. $X \sim F(d_1, d_2)$ is given, for positive values of d_1 and d_2 , by:

$$\varphi_F(x, d_1, d_2) = \begin{cases} \frac{\sqrt{\frac{(d_1 x)^{d_1} (d_2)^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x B(\frac{d_1}{2}, \frac{d_2}{2})}, & \text{for } x \geq 0, \\ 0, & \text{for } x < 0. \end{cases} \tag{2.6}$$

with B the beta function defined by

$$B(a, b) := \int_0^1 a^{t-1} t^{a-1} (1-t)^{b-1} dt. \quad (2.7)$$

The corresponding cumulative distribution function (cdf) is:

$$\Phi_F(x, d_1, d_2) = I_{\frac{d_1 x}{d_1 x + d_2}} \left(\frac{d_1}{2}, \frac{d_2}{2} \right), \quad (2.8)$$

with $I_x(a, b)$ the regularized incomplete beta function defined, for $(a, b) \in \mathbb{R}^2$, as the ratio of the incomplete beta function $B(x; a, b)$ and the complete beta function $B(a, b)$:

$$I_x(a, b) := \frac{B(x; a, b)}{B(a, b)} = \frac{1}{B(a, b)} \int_0^x t^{a-1} (1-t)^{b-1} dt. \quad (2.9)$$

The mode, mean, and variance of the F distribution are listed in Table. 2.2.

	Mode $m[x]$	Mean $\mathbb{E}[x]$	Variance $\text{Var}[x]$
Exists for	$d_1 > 2$ $d_2 > 0$	$d_1 > 0$ $d_2 > 2$	$d_1 > 0$ $d_2 > 4$
Expression	$\frac{d_1-2}{d_1} \frac{d_2}{d_2+2}$	$\frac{d_2}{d_2-2}$	$\frac{2d_1^2(d_2+d_1-2)}{d_2(d_1-2)^2(d_1-4)}$

Table 2.2 – Mode, mean, and variance of the central F law.

Definition 4. Non central F-distribution

A non central F -distribution corresponds to the ratio of two independent random variables $V_1 \sim \chi_{\lambda, d_1}^2$ and $V_2 \sim \chi_{d_2}^2$ with χ_{λ, d_1}^2 the non central χ^2 distribution characterized by the non centrality parameter λ [Abramowitz et al., 1972]

$$\frac{V_1/d_1}{V_2/d_2} \sim F_{\lambda}(d_1, d_2).$$

The pdf of a r.v. $X \sim F_{\lambda}(d_1, d_2)$ is given for positive values of d_1, d_2 , and λ , by:

$$\varphi_{F_{\lambda}}(x, d_1, d_2) = \begin{cases} \sum_{k=0}^{\infty} \frac{e^{-\frac{\lambda}{2}} (\lambda/2)^k}{B(\frac{d_2}{2}, \frac{d_1}{2} + k) k!} \left(\frac{d_1}{d_2} \right)^{\frac{d_1}{2} + k} \left(\frac{d_2}{d_2 + d_1 x} \right)^{\frac{d_1 + d_2}{2} + k} x^{\frac{d_1}{2} - 1 + k}, & \text{for } x \geq 0, \\ 0, & \text{for } x < 0. \end{cases} \quad (2.10)$$

The corresponding cdf is:

$$\Phi_{F_{\lambda}}(x, d_1, d_2) = \sum_{k=0}^{\infty} \left(\frac{(\frac{1}{2}\lambda)^k}{k!} e^{-\frac{\lambda}{2}} \right) I_{\frac{d_1 x}{d_1 x + d_2}} \left(\frac{d_1}{2} + k, \frac{d_2}{2} \right). \quad (2.11)$$

When $\lambda = 0$, obviously the non central F becomes a central F distribution. The mode, mean, and variance of this distribution are listed in Table. 2.3.

	Mode $m[x]$	Mean $\mathbb{E}[x]$	Variance $\text{Var}[x]$
Exists for	$d_1 > 0$ $d_2 > 0$ $\lambda > 0$	$d_1 > 0$ $d_2 > 2$ $\lambda > 0$	$d_1 > 0$ $d_2 > 4$ $\lambda > 0$
Expression	(*)	$\frac{d_2(d_1+\lambda)}{d_1(d_2-2)}$	$2 \frac{(d_1+\lambda)^2 + (d_1+2\lambda)(d_2-2)}{(d_2-2)^2(d_2-4)}$

Table 2.3 – Mode, mean, and variance of the non-central F law. (*) The mode has to be evaluated numerically.

Example 1. Central and non central F distributions

Consider a r.v. $X \sim F(d_1, d_2)$ under the null and $X \sim F_\lambda(d_1, d_2)$ under the alternative. The evolution of the central and non-central F distributions given in (2.6) and (2.10) w.r.t. their parameters are shown in Fig. 2.1 (resp. left and right panels).

One can see the pdf's shape changing with the degrees of freedom (d_1, d_2) and the non centrality parameter (λ) (see legend).

Consider now, for example, the case of the central F density $F(50, 50)$ (yellow curve, left), using the F mean and variance defined in Table. 2.2. One can compare this pdf with a classical Gaussian distribution (black dotted line, left panel) of same mean ($\mu = 1.04$) and variance ($\sigma^2 = 0.09$, cf. Table. 2.2). The same comparison is shown in the right panel comparing the non central F density $F_{70}(50, 50)$ with a Gaussian distribution of same mean ($\mu = 2.50$) and variance ($\sigma^2 = 0.44$, cf. Table. 2.3). In both cases, one can observe large differences between the F distributions and the Gaussians, both in amplitudes (highlighted by the dark regions) and especially wing shapes (highlighted by the yellow regions). If these distributions correspond to φ_T in (2.5), (2.4) and (2.3), they will lead to very different values for the false alarm/detection probabilities at a given threshold. This illustrates the importance of knowing the true distribution of the considered test statistics for accurate false alarm considerations.

Example 2. Distribution of test statistics T

Consider now a test statistic T with $\varphi_T(x|\mathcal{H}_0) = \varphi_F(x, 50, 50)$ under the null, and following a non-central F distribution under the alternative such as $\varphi_T(x|\mathcal{H}_1) = \varphi_{F_{70}}(x, 50, 50)$.

By the analytical expressions of the central (2.6) and non central F -pdf (2.10), one can compute the right and wrong decision probabilities in (2.3), (2.4) and (2.5) for a binary test as in (2.2).

Fig. 2.2 illustrates the decision region plots by representing the density under the null (blue), under the alternative (green) and the probabilities P_M in (2.5) (pink region) and P_{FA} in (2.4) (yellow region) for a given threshold γ . In essence, γ determines how extreme the observed test statistic must be to reject \mathcal{H}_0 . One can understand the critical role of the test's threshold in defining these probability regions. This threshold is always set in a way of making a compromise between the detection and false alarm probabilities.

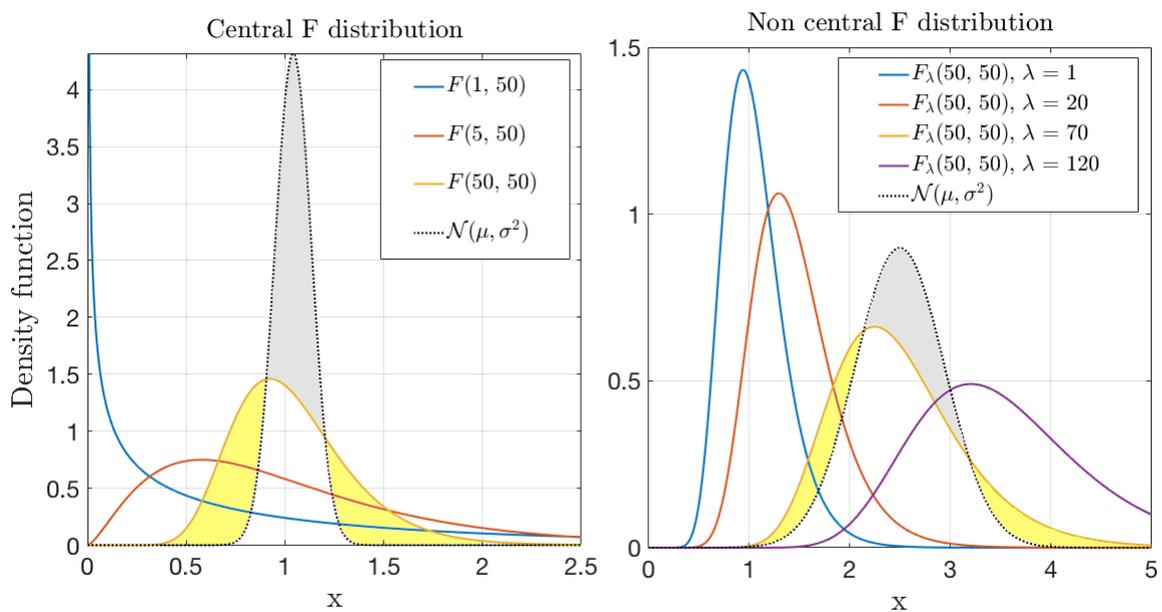


Figure 2.1 – Influence of the parameters of a central (left) and non central distribution (right). In each panel, a Gaussian distribution (black dot curve) is shown with same mean and variance as $F(50, 50)$ and $F_\lambda(50, 50)$.

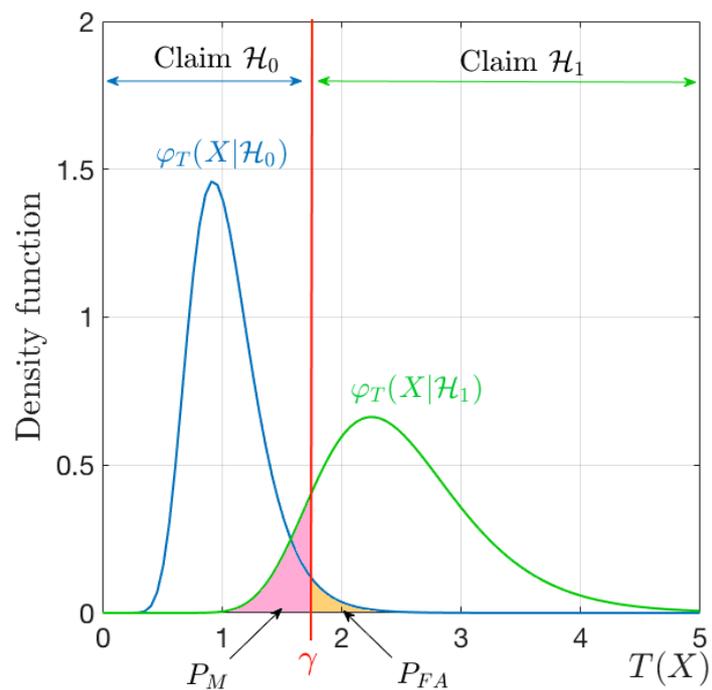


Figure 2.2 – Example of distribution of T under both hypotheses and of the decision regions for a one-sided right tailed test. For a given threshold γ , the probability of missed detection, defined in (2.5), is highlighted in pink while the false alarm probability, defined in (2.4), is highlighted in yellow.

P-value

An important statistical quantity, computed after the data has been collected, is the *P-value*.

Definition 5. P-value

The *P-value*¹ represents the probability that a random variable X is at least as extreme as the observed value x under \mathcal{H}_0 :

$$\begin{aligned} v_X(x) &:= \Pr(X \geq x | \mathcal{H}_0) \\ &= 1 - \Pr(X < x | \mathcal{H}_0) \\ &= 1 - \Phi_X(x), \end{aligned} \tag{2.12}$$

with Φ_X the cdf of X under \mathcal{H}_0 . Consequently, the value v_X is the realization of the random variable $V_X = 1 - \Phi_X(X)$ for $X = x$.

A *P-value* is a probability, and then takes values in interval $[0, 1]$. A small *P-value* may constitute a strong data evidence against \mathcal{H}_0 , but it does not inform on the validity of the alternative. Under \mathcal{H}_0 , by the *Probability Integral Transform theorem*, one can show that the *P-values* are uniformly distributed : $V_X \sim \mathcal{U}(0, 1)$.

Proof of the uniform distribution of V_X under \mathcal{H}_0 . Consider a given threshold $z \in [0, 1]$. The cdf of V_X , say Φ_V , writes

$$\Phi_V(z) := \Pr(V_X < z) = \Pr(1 - \Phi_X(X) < z) = \Pr(\Phi_X(X) > 1 - z).$$

As the cdf is a monotonic, increasing and continuous function, we can write:

$$\Phi_V(z) = \Pr(X > \Phi_X^{-1}(1 - z)) = 1 - \Phi_X(\Phi_X^{-1}(1 - z)) = z. \tag{2.13}$$

Hence, $\forall z \in [0, 1]$, we have $\Phi_V(z) = z$, and $V_X \sim \mathcal{U}(0, 1)$. □

Example 3. Graphical illustration of a P-value

Consider again the case of a test statistic $T \sim F(50, 50)$ under the null.

Fig. 2.3 shows the pdf of T given in (2.6) (blue) with a P_{FA} area (2.4) (yellow region) evaluated for threshold $\gamma = 1.2$. The *P-value* (2.12) evaluated for a given observed value of the test statistic, say $t_{obs} = 1.34$, is shown by the purple dashed region. Here γ corresponds to a P_{FA} in (2.4) of

$$P_{FA} = 1 - \Phi_F(\gamma, 50, 50) = 0.26,$$

and the *P-value* in (2.12) is

$$v_F(t_{obs}) = 1 - \Phi_F(t_{obs}, 50, 50) = 0.15.$$

The black dot curve represents a Gaussian density with same variance and mean as the F-distributed r.v. (cf. Example.1). Comparing the P_{FA} at γ and *P-values* at t_{obs} for the two distributions (blue and black points) in Table. 2.4, one can observe drastic differences and, once again, the importance of knowing the test statistic distribution to control the significance of a detection. Indeed, if we wrongly assume T to be Gaussian, the *P-value* 6.2×10^{-4} will constitute a quite strong evidence against \mathcal{H}_0 . However, if $T \sim F(50, 50)$ under \mathcal{H}_0 , the value of t_{obs} is quite likely (i.e. the true *P-value* is not small).

¹In this manuscript, the *P-values*, usually denoted by the letter P , will be denoted by V because P is used for the periodogram (cf. Sec. 2.2).

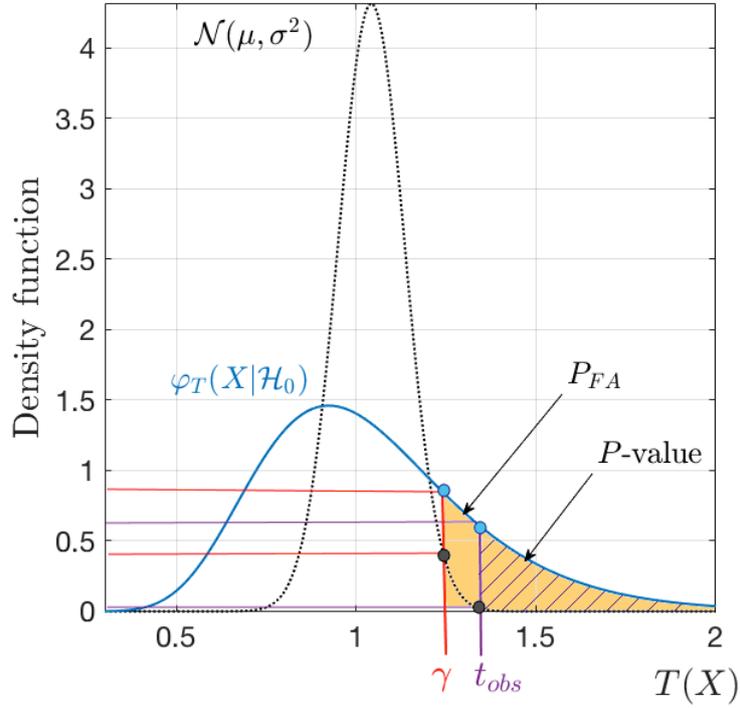


Figure 2.3 – Illustration of the P -value (purple dashed area) for a test statistic T F-distributed (blue). The P_{FA} region (yellow area) and a Gaussian density (black), of same mean, and variance, are shown for comparison.

	F-distribution	Gaussian distribution
$P_{FA}(\gamma)$	2.6×10^{-1}	4.3×10^{-2}
$v_X(t_{obs})$	1.5×10^{-1}	6.2×10^{-4}

Table 2.4 – Numerical difference in terms of P_{FA} (2.4) and P -value (2.12) measured on the F and Gaussian distributions at, respectively, $\gamma = 1.2$ and $t_{obs} = 1.34$.

ROC curve

A *Receiver operation characteristics* (ROC) curve represents the test power (P_{DET}) as function of its size (P_{FA}). In such plots, the two probabilities are determined at a given threshold γ . ROC curves allow to compare the performance of different tests.

Example 4. Evaluation of a ROC curve

Consider again, for test (2.2), the case of a test statistic $T \sim F(50, 50)$ under the null and $T \sim F_\lambda(50, 50)$ under the alternative with $\lambda = 20$ here.

Using the respective cdf of these distributions (resp. Φ_F given in (2.6) and Φ_{F_λ} given in (2.11)), the P_{FA} and P_{DET} writes:

$$\begin{aligned} P_{FA}(\gamma, T) &= 1 - \Phi_F(\gamma, 50, 50), \\ P_{DET}(\gamma, T) &= 1 - \Phi_{F_\lambda}(\gamma, 50, 50). \end{aligned}$$

By extracting γ in the first equation and replacing in the second one, one obtains

$$P_{DET}(P_{FA}, T) = 1 - \Phi_{F_\lambda}(\Phi_F^{-1}(1 - P_{FA}, d_1, d_2), 50, 50),$$

with Φ_F^{-1} the inverse F cdf.

This allows to compute ROC curves shown in Fig. 2.4 (red solid line). The diagonal illustrates the random line, i.e. the detection behavior of a test that will not take any advantages of the observed data and will give the same FA and detection probabilities what ever the considered threshold. A ROC curve observed on such diagonal represents the worst case scenario ($P_{FA} = P_{DET}$). As the threshold γ increases, both probabilities decreases, with two limiting cases: when $\gamma \rightarrow \infty$, $P_{FA} = P_{DET} = 0$ and when $\gamma \rightarrow 0$, $P_{FA} = P_{DET} = 1$ (for positive test statistics). The dotted red line represents a ROC curve of a more powerful test and the dashed red line of a less powerful one. The *Area Under the Curve* (AUC, yellow region) can also be used to measure the global test power over all P_{FA} . This area is by construction in the interval $]0.5, 1]$, with $AUC = 0.5$ for the worst case (as the ROC is on the diagonal) and $AUC = 1$ for the ideal case (with $P_{DET} = 1, \forall P_{FA}$).

2.1.3 Frequentist approaches: tests based on likelihood functions

One of the main stream for building tests is based on the likelihood functions for comparing and scoring both hypotheses.

The likelihood

In statistics, the likelihood is defined as a conditional probability function of the observations, conditioned to the parameters of a given model. The likelihood of N random variables $\mathbf{X} = [X_1, \dots, X_N]^T$ for parameter $\boldsymbol{\theta}$ is the joint density of these N r.v.:

$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{x}) := \varphi_{\mathbf{X}}(\mathbf{x}, \boldsymbol{\theta}) = \varphi_{X_1, \dots, X_N}(\mathbf{x}, \boldsymbol{\theta}). \quad (2.14)$$

If the X_i are independent, the likelihood function is given by the product of these densities:

$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{x}) = \prod_{i=1}^N \varphi_{X_i}(x_i, \boldsymbol{\theta}). \quad (2.15)$$

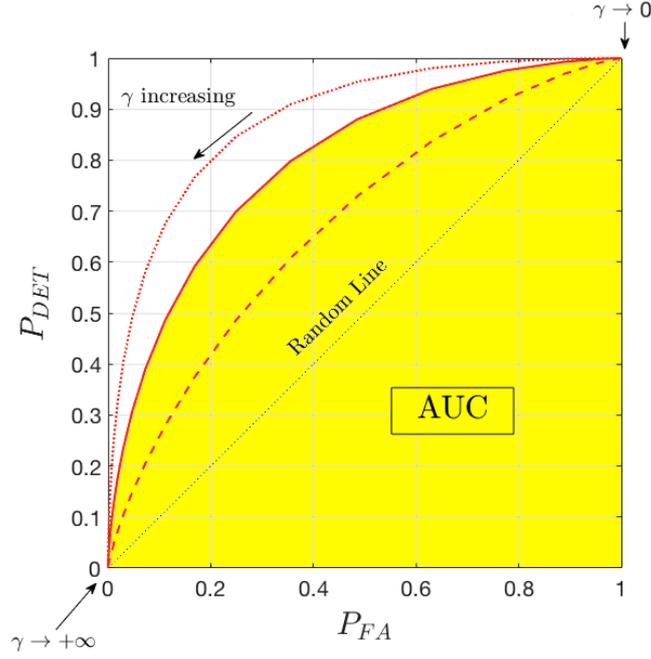


Figure 2.4 – Illustration of a test performance through the ROC curve and AUC.

The higher the likelihood, the better the model’s parameters fit with the observed data. Maximizing (2.15) with respect to θ leads to the maximum likelihood estimate (MLE):

$$\hat{\theta}_{ML} := \arg \max_{\theta} \mathcal{L}(\theta, \mathbf{x}). \quad (2.16)$$

The “hat” over θ indicates that this value is an estimate. This general notation will be used for all estimates in this document.

If the parameter vector θ is known under both hypotheses (θ_0 and θ_1) the *Likelihood Ratio* (LR) test consists in comparing to a threshold the ratio of the two likelihoods evaluated under both hypotheses:

$$T_{LR}(\mathbf{x}) = \frac{\varphi_{\mathbf{X}}(\mathbf{x}, \theta_1)}{\varphi_{\mathbf{X}}(\mathbf{x}, \theta_0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\geq}} \gamma. \quad (2.17)$$

According to the Neyman-Pearson lemma, in the case of a binary hypothesis where θ is known under both hypotheses, the LR test is the most powerful detection test (i.e. no test has larger P_{DET} at a given P_{FA} level) [Kay, 1998].

Lemma 1 (Neyman-Pearson lemma). *In order to maximize the P_{DET} for a fixed $P_{FA} = \alpha$, we reject \mathcal{H}_0 if $T_{LR}(\mathbf{x}) > \gamma$ where*

$$P_{FA} = \Pr(T_{LR}(\mathbf{x}) > \gamma | \mathcal{H}_0) = \alpha. \quad (2.18)$$

If θ is unknown under one (or both) hypotheses, the *Generalized Likelihood Ratio* (GLR) test consists in computing the same ratio as in (2.17), but with θ replaced by the MLE ($\hat{\theta}_{ML}$):

$$T_{GLR}(\mathbf{x}) = \frac{\sup_{\hat{\theta}_{ML} \in \Theta_1} (\varphi_{\mathbf{X}}(\mathbf{x}, \hat{\theta}_{ML}))}{\sup_{\hat{\theta}_{ML} \in \Theta_0} (\varphi_{\mathbf{X}}(\mathbf{x}, \hat{\theta}_{ML}))} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\geq}} \gamma. \quad (2.19)$$

with (Θ_0, Θ_1) the respective parameter spaces under the two hypotheses.

In Sec. 2.4, we will study other detection tests, which are related to the GLR test (2.19) under some conditions of the hypothesis testing (e.g. the Fisher’s test in (2.58)) and also other determinist approaches (e.g. the Higher Criticism in (2.70) or the Berk Jones tests in (2.74)).

2.1.4 A brief comment on Bayesian approaches

A second main stream of detection approaches exists (but we will not focus on it in this manuscript) namely, the Bayesian approach.

Bayesian methods are an alternative to the frequentist approach. They consider the parameters vector $\boldsymbol{\theta}$ as random variables. These methods are based on the *Bayes' theorem* which, using a prior distribution for $\boldsymbol{\theta}$ and a conditional distribution of the data, allows to derive a posterior probability for the chosen model according to the observed data. A Bayesian analysis test consists in comparing the probability of the model according to the a posteriori pdf (see [Wakefield, 2013] for a recent review).

In extrasolar planet detection, a series of Bayesian techniques exist to search for Keplerian planetary signatures of the form of (1.3) (e.g. [Ford and Gregory, 2007, Balan and Lahav, 2009]) in different kind of noises (e.g. White Gaussian noise (WGN) or Gaussian processes [Haywood, 2015]).

2.2 Periodogram(s)

Consider the general model hypotheses

$$\begin{cases} \mathcal{H}_0 : X(t) = E(t) \\ \mathcal{H}_1 : X(t) = R(t) + E(t) \end{cases} \quad (2.20)$$

with $E(t)$ a stationary Gaussian noise and $R(t)$ the (possibly quasi-) periodic signature of interest.

Before deriving hypotheses tests for model (2.20), let us look at the representation of $R(t)$ in the frequency domain. This leads to introduce the *periodogram*.

2.2.1 The classical periodogram

Consider a discrete time series $\mathbf{x}(t) = [x(t_1), \dots, x(t_N)]^\top$ with a regular sampling step $\Delta t = t_{j+1} - t_j$ defined for $\Delta t \in \mathbb{R}^{+*}$. The Discrete Fourier transform (DFT) of \mathbf{x} , at frequency ν , is

$$\text{DFT}(\mathbf{x}) := \frac{1}{N} \langle \mathbf{g}(\nu), \mathbf{x} \rangle := \frac{1}{N} \sum_{j=1}^N \mathbf{x}(t_j) e^{-i2\pi t_j \nu}.$$

with $\langle \cdot, \cdot \rangle$ the scalar product and

$$\mathbf{g}(\nu) := [e^{i2\pi\nu}, \dots, e^{i2\pi\nu t_N}]^\top. \quad (2.21)$$

The considered frequencies ν are *Fourier frequencies*:

$$\nu_k := \frac{k}{N\Delta t}, \quad \text{for } k = 0, \dots, N-1. \quad (2.22)$$

One can show that the set $\mathbf{g}(\nu_k)$ forms an orthogonal basis of vectors². Defining a matrix $\mathbf{G} := [\mathbf{g}(\nu_0), \dots, \mathbf{g}(\nu_{N-1})]^\top$, the complete DFT vector (containing all frequencies) writes

$$\text{DFT}(\mathbf{x}) := \frac{1}{N} \mathbf{G}\mathbf{x}. \quad (2.23)$$

As \mathbf{G} is composed of N basis vectors, any signal \mathbf{x} can be decomposed on this basis.

²i.e., $\langle \mathbf{g}(\nu_k), \mathbf{g}(\nu_{k'}) \rangle = 0$ if $k \neq k'$ and N otherwise.

Fourier analysis for periodicity detection dates back to Schuster, who introduced the notion of *periodogram* as a measure for the strength of the components of the DFT vector in (2.23) [Schuster, 1898]:

$$\begin{aligned}
P(\nu) &= \frac{1}{N} \left| \sum_{j=1}^N \mathbf{x}(t_j) \mathbf{g}^H \right|^2 \\
&= \frac{1}{N} \left| \sum_{j=1}^N x(t_j) e^{-i2\pi\nu t_j} \right|^2 \\
&= \frac{1}{N} \left| \sum_{j=1}^N x(t_j) \left(\cos(2\pi\nu t_j) + i \sin(2\pi\nu t_j) \right) \right|^2.
\end{aligned} \tag{2.24}$$

The sine and cosine terms of the complex exponential in (2.24) are orthogonal for all $k \neq k'$. It follows that $P(\nu_k)$ and $P(\nu_{k'})$ are independent for $k \neq k'$ for any second order stationary Gaussian process with autocorrelation function $r_E \xrightarrow{N \rightarrow +\infty} 0$ (cf. Theorems 5.2.1 and 5.2.4 of [Brillinger, 1981]). When the data $x(t_j)$ are uncorrelated, half of this subset of ordinates $P(\nu_k)$ are independent ($\frac{N}{2}$ frequencies if N is even and $\frac{N-1}{2}$ if N is odd). This is also asymptotically true for any Gaussian noise [Quinn and Hannan, 2001]. In the following, for simplicity and without loss of generality, we assume N even and define the subset of $(\frac{N}{2} - 1)$ Fourier frequencies corresponding to $k \in \Omega := \{1, \dots, \frac{N}{2} - 1\}$.

Under several assumptions, the statistical properties of the periodogram allow to characterize P in (2.24) as an estimate of the power spectral density (PSD). These assumptions are:

- (i) The sampling is regular;
- (ii) The periodogram is evaluated at Fourier frequencies $(\nu_k, (2.22))$;
- (iii) The number of samples N is large (asymptotic regime);
- (iv) Under \mathcal{H}_0 , the random process X is stationary Gaussian with known statistics.

Under these assumptions (i)–(iv), any finite set of periodogram ordinates near some fixed frequency is asymptotically independent and exponentially distributed with parameters dependent on the noise PSD at the relevant frequencies (this assumption actually extends to linear processes other than Gaussian, see Theorem 4 of [Quinn and Hannan, 2001]).

In detection, this property is extensively used to obtain the probabilities of false alarm of tests based on periodogram ordinates (as in (2.4)). If under the alternative of model (2.20) (periodic signal in noise) the distribution of P can also be precisely characterized, these properties translate into an accurate control of the detection probability (cf. (2.3)) as well.

Despite a large literature on the subject (e.g. [Bartlett, 1950, Grenander and Rosenblatt, 1957, Brillinger, 1981, Priestley, 1981, Brockwell and Davis, 1991, Bloomfield, 2000, Quinn and Hannan, 2001, Stoica and Moses, 2005]), the detection of periodic signals remains an active field of research because in practical situations some (or all) assumptions (i)–(iv) above may not be met [Li, 2014]. The following paragraphs summarize useful properties of the periodogram.

2.2.2 Periodogram of a single sinusoid

It is useful for the following to look at the periodogram structure on the example of an evenly sampled pure sinusoidal signal $R(t) = \alpha_s \sin(2\pi f_s t + \varphi_s)$ in model (2.20), with $\alpha_s \in \mathbb{R}^+$ the signal amplitude, $f_s \in \mathbb{R}$ its frequency and $\varphi_s \in [0, 2\pi]$ a constant phase term.

When evaluated at an arbitrary frequency ν and using $1/i = e^{-i\pi/2}$, the periodogram writes:

$$\begin{aligned}
P(\nu) &= \frac{1}{N} \left| \sum_{j=1}^N \alpha_s \sin(2\pi f_s t_j + \varphi_s) e^{-i2\pi\nu j \Delta t} \right|^2 \\
&= \frac{1}{N} \left| \sum_{j=1}^N \alpha_s \frac{e^{i2\pi f_s t_j + \varphi_s} - e^{-i2\pi f_s t_j + \varphi_s}}{2i} e^{-i2\pi\nu j \Delta t} \right|^2 \\
&= \frac{\alpha_s^2}{4N} \left| \sum_{j=1}^N e^{i2\pi(f_s - \nu)t_j} e^{i\varphi_s - i\pi/2} - e^{-i2\pi(f_s + \nu)t_j} e^{-i\varphi_s - i\pi/2} \right|^2.
\end{aligned} \tag{2.25}$$

To develop (2.25), let us define the Dirichlet and Fejèr kernels.

Definition 6. Dirichlet kernel

The Dirichlet kernel is related to the discrete spectral window of the observations and defined as [Li, 2014]:

$$\begin{aligned}
D_N(\nu) &= \frac{1}{N} \sum_{n=1}^N \sum_{j=1}^N \delta(t - j\Delta t) e^{-i2\pi\nu j \Delta t} \\
&= \frac{1}{N} \sum_{j=1}^N e^{i2\pi\nu j \Delta t} \\
&= \frac{1}{N} \sum_{j=1}^N \left(e^{i2\pi\nu \Delta t} \right)^j \quad (\text{sum of a geometric series}) \\
&= \frac{1}{N} \left(\frac{1 - \left(e^{i2\pi\nu \Delta t} \right)^N}{1 - \left(e^{i2\pi\nu \Delta t} \right)} \right) \\
&= e^{-i\pi\nu(N+1)\Delta t} \left(\frac{\sin(\pi\nu N \Delta t)}{N \sin(\pi\nu \Delta t)} \right).
\end{aligned} \tag{2.26}$$

Kernel D_N is a cardinal sine like function, with maximum value at $\nu_k = 0$ and zeros at the Fourier frequencies $\nu_k \neq 0$.

The left panel of Fig. 2.5 illustrates the modulus of the Dirichlet kernel as a function of ν for different values of the time series length (N). This kernel tends to a Dirac impulsion $\delta(t)$ for infinite N .

Definition 7. Fejèr kernel

The squared modulus of D_N is called the Fejèr kernel [Li, 2014]:

$$K_N(\nu) := \left| D_N(\nu) \right|^2 = \left| \frac{1}{N} \sum_{j=1}^N e^{i2\pi\nu j} \right|^2 = \left(\frac{\sin(N\pi\nu)}{N \sin(\pi\nu)} \right)^2. \tag{2.27}$$

The periodogram of the pure sine signal, in (2.25), can then be written as a function of these kernels

$$\begin{aligned}
P(\nu) &= \frac{\alpha_s^2}{4N} \left| N D_N(f_s - \nu) e^{i(\varphi_s - \pi/2)} + N D_N(f_s + \nu) e^{-i(\varphi_s + \pi/2)} \right|^2, \\
&= \frac{N\alpha_s^2}{4} \left[K_N(f_s - \nu) + K_N(f_s + \nu) + 2\sqrt{K_N(f_s - \nu)K_N(f_s + \nu)} \cos(2\pi f_s(N + 1) + 2\varphi_s) \right].
\end{aligned} \tag{2.28}$$

Written as in (2.28), one can see that the periodogram of a pure sine is composed by two superimposed Fejèr kernels centred at opposite frequencies (f_s and $-f_s$) plus an “interference” term.

The right panel of Fig. 2.5 shows the periodogram of a sinusoidal signal ($\alpha_s = 2, \varphi_s = 0$) with a Fourier frequency ($f_s = \nu_{451}$, green), and with a frequency between two Fourier frequencies (red). The superposition of the Fejèr kernel evaluated for continuous frequencies³, ν , is shown by the black solid line.

- If the frequency f_s is on the Fourier frequency grid (i.e., $f_s = \nu_k$, for some k), the periodogram achieves its maximum value in $\pm f_s$ and is cancelled at the other frequencies. Using (2.27), the amplitudes of the periodogram at $\pm f_s$ are $\frac{N\alpha_s^2}{4}$ as $K_N(0) = 1$ and the interference term in (2.28) is cancelled. In such case, the Fejèr kernels are sampled on their global maxima in $\pm f_s$ and the periodogram crosses its zeros at all ν_k .
- If f_s is not on the Fourier grid, the sinusoidal component is orthogonal to none of the complex exponentials sampled at Fourier frequencies (the different ν_k in (2.22)). The kernels K_N and interference terms in (2.28) are sampled at values where they are not null and the periodogram has many (all, in fact) non zeros values at $\nu_k \neq \pm f_s$. In this case, the Fejèr kernel illustrates the effect of *spectral leakage*: the “power of the sinusoid with non-Fourier frequency” spreads over all frequencies.

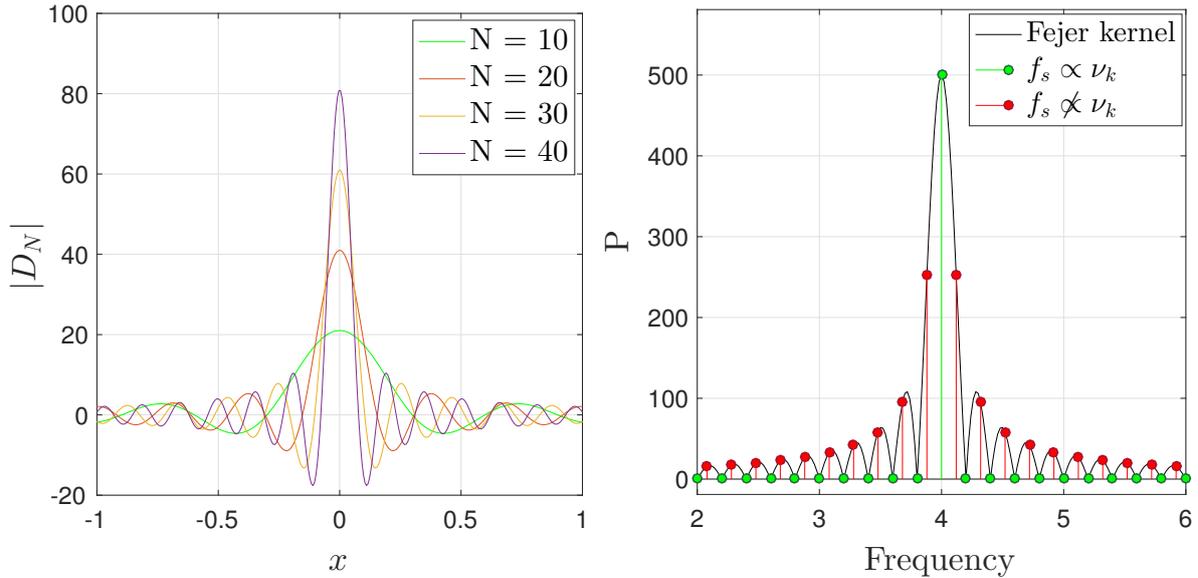


Figure 2.5 – Left: Illustration of the Dirichlet kernel defined in (2.26). Right: Illustration of the spectral leakage effect on a sinusoidal signal. Comparison of the case where the Fourier frequencies fall on the signal frequency f_s (green stems) or not (red stems). The solid black curve represents the Fejèr kernel $K_N(\nu - f_s)$, defined in (2.27), centred on f_s .

We note that for a complex signal of the form $R(t) = \alpha e^{2\pi i f_s t}$, the periodogram expression (2.24) is simpler and becomes:

$$P(\nu) = \frac{1}{N} \left| \sum_{j=1}^N R(t_j) e^{-2\pi i t_j \nu} \right|^2 = \frac{\alpha^2}{N} \left| \sum_{j=1}^N e^{2\pi i t_j (f_s - \nu)} \right|^2 = \frac{\alpha^2}{N} \left| N D_N(f_s - \nu) \right|^2 = \alpha^2 N K_N(f_s - \nu).$$

³We used an oversampling factor of 10 to mimic a continuous vector of frequencies.

2.2.3 Parseval's identity

Using the Parseval's identity, one can show that the periodogram is related to the sample variance. Here is the demonstration.

Consider N even and a time series with standard deviation σ_X .

(i) Using Fourier series to develop $X(t)$, the empirical variance can be written as a sum of sines and cosines functions

$$\begin{aligned}\hat{\sigma}_X^2 &:= \frac{1}{N} \sum_{t=1}^N (X(t) - \bar{X})^2 \\ &= \frac{1}{N} \sum_{t=1}^N \left(\left[\sum_{k=0}^K A_k \cos(2\pi\nu_k t) + B_k \sin(2\pi\nu_k t) \right] - \bar{X} \right)^2 \\ &= \frac{1}{N} \sum_{t=1}^N \left(\sum_{k=1}^K A_k \cos(2\pi\nu_k t) + B_k \sin(2\pi\nu_k t) \right)^2\end{aligned}\tag{2.29}$$

with K the number of considered components, and $\bar{X} = A_0 = \frac{1}{N} \sum_{t=1}^N X(t)$ the sample mean.

By using the orthogonality relations between sines and cosines at Fourier frequencies ν_k , $\forall k \in \Omega := \{1, \dots, \frac{N}{2} - 1\}$, one can show that:

$$\begin{cases} A_k = \frac{2}{N} \sum_{t=1}^N X(t) \cos(2\pi\nu_k t), \\ B_k = \frac{2}{N} \sum_{t=1}^N X(t) \sin(2\pi\nu_k t). \end{cases}\tag{2.30}$$

Consequently, the variance estimate (2.29) writes

$$\hat{\sigma}_X^2 = \frac{1}{2} \sum_{k=1}^{\frac{N}{2}-1} (A_k^2 + B_k^2).\tag{2.31}$$

(ii) The periodogram of $X(t) - \bar{X}$ writes :

$$P(\nu_k) = \frac{1}{N} \left(\left[\sum_{t=1}^N (X(t) - \bar{X}) \cos(2\pi\nu_k t) \right]^2 + \left[\sum_{t=1}^N (X(t) - \bar{X}) \sin(2\pi\nu_k t) \right]^2 \right).$$

Using (2.30) and considering Fourier frequencies ν_k with $\forall k \in \Omega$, the periodogram can be written as:

$$P(\nu_k) = \frac{N}{4} (A_k^2 + B_k^2).\tag{2.32}$$

(iii) The Parseval's identity follows by combining (2.31) and (2.32)

$$\hat{\sigma}_X^2 = \frac{2}{N} \sum_{k=1}^{N/2} P(\nu_k).\tag{2.33}$$

Hence, the total dispersion of the time series is a scaled sum of the power at each frequency $P(\nu_k)$. In other words, each ordinate $P(\nu_k)$ is homogenous to a variance and so to an estimate of the signal PSD (here σ_X^2).

The Parseval's identity is also consistent with the spectral leakage: because of (2.33), the energy of a pure sine at non Fourier frequency is not only concentrated on $\nu_k = f_s$ but spreads over the neighbourhood components (cf. Fig. 2.5).

2.2.4 Periodogram resolution

The resolution of the periodogram, i.e. its capacity to distinguish between two distinct frequencies, can only be increased by adding new observed data (increasing N , see Fig. 2.5). If the number of observed data is fixed, one can apply the technique of *zero padding*, which consists in adding zeros at the end of the observed time series ($N' > N$) to work with a finer Fourier frequency grid $\{\frac{k}{N'}\}$. The zero padding technique does not improve the spectral resolution. When $N' \rightarrow \infty$, the resulting periodogram is called *continuous periodogram* (see black curve in Fig. 2.5).

2.2.5 Bias and variance of the periodogram

Consider an evenly sampled time series \mathbf{X} from a random process X with PSD S_X estimated by the periodogram $\mathbf{P} := [\mathbf{P}_{\nu_1}, \dots, \mathbf{P}_{\nu_N}]^\top$. One wants to quantifiante how \mathbf{P} deviates, in average, to the true PSD S_X (i.e. the estimation bias) as well as the variability of \mathbf{P} from one training data sample \mathbf{X} to another (i.e. the estimation variance).

Definition 8. Bias and variance

The bias b and variance of an estimator $\hat{\theta}$ of some scalar parameter θ are defined as:

$$b[\hat{\theta}] := \mathbb{E}[\hat{\theta}] - \theta$$

$$\text{Var}[\hat{\theta}] := \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)^2\right].$$

In the non asymptotic regime ($N \ll \infty$), one can show that the periodogram is a biased PSD estimate of S_X as $\mathbb{E}[P(\nu_k)] \neq S_X(\nu_k)$. However, the periodogram is asymptotically an unbiased and inconsistent estimate (cf. [Brillinger, 1981], Chap. 5, theorems 5.2.1 and 5.2.4):

$$\begin{aligned} \mathbb{E}[P(\nu_k)] &= S_X(\nu_k) + \mathcal{O}(1/N), \\ \text{Var}[P(\nu_k)] &= S_X(\nu_k)^2 + \mathcal{O}(1/N), \\ \text{Cov}[P(\nu_k), P(\nu_{k'})] &= \mathcal{O}(1/N), \quad \forall k \neq k'. \end{aligned} \tag{2.34}$$

with \mathcal{O} the Bachmann-Landau notation⁴, and Cov the covariance function of P

$$\text{Cov}[P(\nu_k), P(\nu_{k'})] := \mathbb{E}\left[\left(P(\nu_k) - \mathbb{E}[P(\nu_k)]\right)\left(P(\nu_{k'}) - \mathbb{E}[P(\nu_{k'})]\right)\right].$$

2.2.6 Asymptotic distribution in the case of a pure random process

The asymptotic distribution of the periodogram under the null hypothesis of model (2.20) is given in Theorem. 6.4 of [Li, 2014] (p.178).

⁴We note a function $f(x)$ as $f(x) = \mathcal{O}(g(x))$ if $f(x)$ converges to a function $g(x)$ (up to a constant) as x goes to infinity.

Theorem 1. (Asymptotic distribution of P in the case of noise - Theorem. 6.4 of [Li, 2014]). Consider a time series $X(t) = E(t)$ as a sampled zero-mean stationary complex Gaussian process with autocorrelation function such as $\sum_k |r_E(k)| < \infty$ and strictly positive PSD $S_E > 0$. In this case, the asymptotic distribution of the periodogram under the null hypothesis of model (2.20) is:

$$P(\nu_k | \mathcal{H}_0) \sim \begin{cases} \frac{1}{\rho(\nu_k)} \frac{S_E(\nu_k)}{2} \chi_2^2, & \forall k \in \Omega, \\ \frac{1}{\rho(\nu_k)} S_E(\nu_k) \chi_1^2, & \text{for } k = 0, \frac{N}{2}. \end{cases} \quad (2.35)$$

where, $\forall k \in \Omega$:

$$\rho_k := \rho(\nu_k) = 1 + \mathcal{O}(r_N). \quad (2.36)$$

with

$$r_N := \sum_{u=-N}^N \min\left(1, \frac{|u|}{N}\right) |r_E(u)| \xrightarrow{N \rightarrow +\infty} 0. \quad (2.37)$$

(cf. Appendix. B). Moreover, with m fixed as $N \rightarrow \infty$, the r.v. in $\{P(\nu_k) : \nu_k \in \Omega_m\}$ are asymptotically independent for any m -subset $\Omega_m \in \Omega$. When r_N is small, the distribution of $P(\nu_k | \mathcal{H}_0)$ for any given $\nu_k \in \Omega$ can be approximated by $\frac{1}{2} S_E(\nu_k) \chi_2^2$.

As seen in (2.34), unlike the case of WGN, the periodogram ordinates are no longer independent for finite sample sizes when the noise is colored. Theorems 5.2.1 and 5.2.4 of [Brillinger, 1981] show nevertheless that the periodogram ordinates are asymptotically uncorrelated as $N \rightarrow \infty$. In practice of course, N is finite. However, the asymptotic results above are indeed useful in many practical cases, because the independence of the ordinates of P at different frequencies and the asymptotic distribution $\frac{1}{2} S_E(\nu_k) \chi_2^2$ often hold with good accuracy if the correlation structure decays fast enough (these assumptions are made in Whittle's likelihood for instance, cf. Remark 6.6, p.178 of [Li, 2014]). Our derivations below will be based on approximate independence assumption, which will be checked against simulations in Chap. 3 (e.g., cf. Sec. 3.5.1).

In practice, the statistics of the noise are often not (or only partially) known, and so are those of the P ordinates (since S_E is involved in (2.35)). In this case, it is thus very difficult to assess how reliable is the claimed false alarm rate. This important point will be the subject of Sec. 2.4.2 and 2.6.

2.2.7 Asymptotic distribution of P in the case of a periodic signal in noise

Let us consider a time series $X(t)$ composed of a sum of N_S complex sines and a stationary colored Gaussian noise $E(t)$:

$$X(t) = \sum_{j=1}^{N_S} \alpha_j e^{2\pi i f_j t + \varphi_j} + E(t) \quad (2.38)$$

Denote by $\boldsymbol{\theta} := [(\alpha_1, f_1, \varphi_1), \dots, (\alpha_{N_S}, f_{N_S}, \varphi_{N_S})]^\top$ the signal parameters for amplitudes, frequencies (not restricted to the Fourier grid) and phases.

The asymptotic distribution of the periodogram under the alternative hypothesis of model (2.20) is given in Corollary 6.2.b of [Li, 2014] (p.196).

Theorem 2. (Asymptotic distribution of P in the case of periodic signals in noise - Corollary 6.2.b of [Li, 2014]). *The asymptotic distribution of P under \mathcal{H}_1 for a model as in (2.38) is*

$$P(\nu_k|\mathcal{H}_1) \sim \begin{cases} \frac{1}{\rho(\nu_k)} \frac{S_E(\nu_k)}{2} \chi_{\lambda_k, 2}^2, & \forall k \in \Omega, \\ \frac{1}{\rho(\nu_k)} S_E(\nu_k) \chi_{\lambda_k, 1}^2, & \text{for } k = 0, \frac{N}{2} \end{cases} \quad (2.39)$$

with ρ_k given in (2.36) and λ_k the non centrality parameter associated to the χ_2^2 distribution given by

$$\lambda_k := \lambda(\nu_k) = 2\rho(\nu_k)N \frac{1}{S_E(\nu_k)} \left| \sum_{j=1}^{N_S} \alpha_j D_N(\nu_k - f_j) \right|^2 \quad (2.40)$$

with D_N defined in (2.26).

The case of a real sinusoidal signal under \mathcal{H}_1 instead of the complex signal in (2.38) will be developed in Chap. 3.

Example 5. Distribution of the periodogram: case of a White Gaussian noise

Consider under \mathcal{H}_0 a time series involving a WGN $X_i = X(t_i) \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, $\forall i$ and under \mathcal{H}_1 a complex sine of the form of (2.38) in this noise ($\alpha_s = 0.5$, $f_s = 0.1$, $\varphi_s = 0$).

As $\mathbf{X} := [X_1, \dots, X_N]^\top$ is random, the periodogram (2.24) is also random. It is easy and instructing to retrieve the results of the two theorems above in this simple case and in the case of a colored noise in the next example.

Under \mathcal{H}_0 , the periodogram at frequencies ν_k ($\forall k \in \Omega$) can be decomposed as:

$$\begin{aligned} P(\nu_k|\mathcal{H}_0) &= \frac{1}{N} \left| \sum_{j=1}^N X(t_j) \left(\cos(2\pi\nu_k t_j) + i \sin(2\pi\nu_k t_j) \right) \right|^2 \\ &= \frac{1}{N} \left(\sum_{j=1}^N X(t_j) \cos(2\pi\nu_k t_j) \right)^2 + \frac{1}{N} \left(\sum_{j=1}^N X(t_j) \sin(2\pi\nu_k t_j) \right)^2 \\ &= C(\nu_k)^2 + D(\nu_k)^2. \end{aligned} \quad (2.41)$$

Terms $C(\nu_k)$ and $D(\nu_k)$ are two linear combinations of independent normal variables. Thanks to Parsen's theorem, one can deduce (Theorem 4, p.90 [Parsen, 1962])

$$\begin{cases} C(\nu_k) \sim \chi_1^2 & \text{and } D(\nu_k) \sim \chi_1^2 & \Rightarrow & P(\nu_k|\mathcal{H}_0) \sim \chi_1^2 + \chi_1^2 = \chi_2^2, & \forall k \in \Omega \\ C(\nu_k) \sim \chi_1^2 & \text{and } D(\nu_k) = 0 & \Rightarrow & P(\nu_k|\mathcal{H}_0) \sim \chi_1^2, & \text{for } k = \left\{0, \frac{N}{2}\right\}. \end{cases}$$

It follows that, $P(\nu_k)$ is exactly distributed under the null as

$$\frac{2}{\sigma^2} P(\nu_k|\mathcal{H}_0) \sim \chi_2^2 \quad \text{for } k \in \Omega. \quad (2.42)$$

Following the same reasoning under \mathcal{H}_1 for a complex signal in WGN, using (2.39) the distribution under the alternative is

$$\frac{2}{\sigma^2} P(\nu_k|\mathcal{H}_1) \sim \chi_{2, \lambda_k}^2, \quad \text{for } k \in \Omega. \quad (2.43)$$

with the non centrality parameter in (2.40) given by

$$\lambda(\nu_k) = \begin{cases} 0 & \nu_k \neq f_s \\ \frac{2N\alpha_s^2}{\sigma^2} & \nu_k = f_s. \end{cases} \quad (2.44)$$

The left panel of Fig. 2.6 illustrates the periodogram distribution under \mathcal{H}_0 at one fixed Fourier frequency ($\nu_k = \nu_{61} = 0.1$, $\Delta t = 1$) for $X_i = E_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$. The empirical pdf has been computed with 10^5 Monte Carlo simulations (normalized histogram) with times series of length $N = 100$. The theoretical distribution $\varphi_{\chi_2^2}(x) = e^{-x/2}$ is over-plotted in red. The right panel illustrates the distribution of $P(\nu_{61})$ under the alternative for a model of a WGN and a single complex signal of the form (2.38) with amplitude $\alpha_s = 0.5$, frequency on grid $f_s = \nu_{61}$ and zero phase. According to (2.43), the theoretical distribution is a non central χ_2^2 with non centrality parameter (2.44) $\lambda = \frac{2N\alpha_s^2}{\sigma^2} = 50$ (red curve).

To illustrate the influence of N and the signal amplitude α_s in this distribution, we also represent the theoretical pdf for resp. $N = 40$ (i.e. $\lambda = 20$) and $\alpha_s = 0.75$ ($\lambda = 112.5$) by the green and yellow curves. One can see for a large number of data points (or a large signal amplitude) the distribution leads to higher periodogram values under \mathcal{H}_1 , which in turn lead to an easier detection.

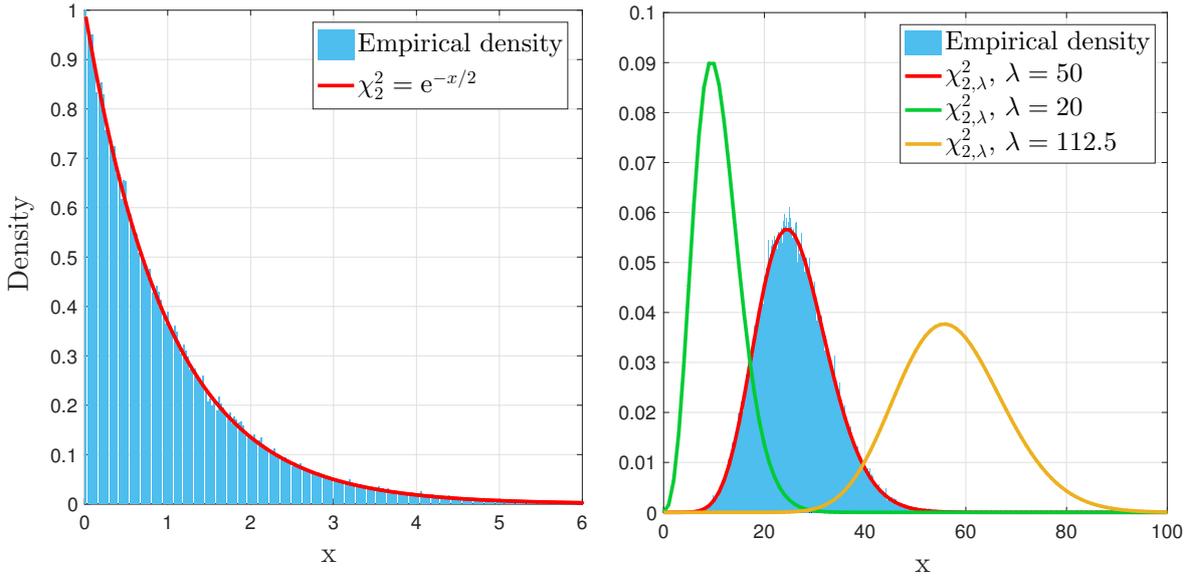


Figure 2.6 – Periodogram distributions at one Fourier frequency (ν_{61}) under the null (left) and the alternative (right) of the model of Example. 6. Blue histograms represent the empirical distributions obtained through 10^5 MC simulations of a WGN $\stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ under \mathcal{H}_0 , and with a complex sine function in this noise with frequency on grid $f_s = \nu_{61}$ under \mathcal{H}_1 . Red curves are the theoretical density derived in (2.35) and (2.39). In the right panel, three theoretical distributions are shown corresponding resp. to $(N, \alpha_s) = (40, 0.5)$, $(100, 0.5)$, $(100, 0.75)$ for the green, red and yellow curves.

Example 6. Periodogram at the signal frequency in colored noise

Consider a complex signal of the form of (2.38) without noise. When the signal frequency f_s is on grid, the periodogram value in $\nu_k = f_s$ is given by (2.25) :

$$P(\nu_k) = N \frac{\alpha_s^2}{2}.$$

When a noise $E(t)$ of PSD S_E is present in (2.38), using the periodogram distribution under \mathcal{H}_1 in (2.39) and the properties of the non central χ^2 distribution (cf. Table. A.3, Appendix. A), one can deduce for $\nu_k = f_s$:

$$\begin{cases} \mathbb{E}[P(\nu_k)] = \frac{(2 + \lambda_k)S_E(\nu_k)}{2} = (1 + \frac{\lambda_k}{2})S_E(\nu_k), \\ \text{Var}[P(\nu_k)] = \frac{2(2 + 2\lambda_k)S_E^2(\nu_k)}{2^2} = (1 + \lambda_k)S_E^2(\nu_k). \end{cases}$$

At the other frequencies $\nu_k \neq f_s$, the non centrality parameter is null and, following the periodogram distribution given in (2.35), the periodogram esperance and variance are

$$\begin{cases} \mathbb{E}[P(\nu_k)] = S_E(\nu_k), \\ \text{Var}[P(\nu_k)] = S_E(\nu_k)^2. \end{cases}$$

Consequently, at the signal frequency, the periodogram values increase with N and α^2 , and they do not change at the other frequencies.

Example 7. Periodogram of a WGN: bias and variance

Consider a White Gaussian noise $X(t) \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$. In this case, as seen in the previous example, the periodogram ordinates evaluated at ν_k are χ_2^2 distributed (2.42). Using the Esperance and variance of a χ^2 distribution (cf. Table. A.1, Appendix. A), it follows that

$$\begin{aligned} \mathbb{E}[P(\nu_k)] &= \mathbb{E}\left[\frac{\sigma^2 \chi_2^2}{2}\right] = \frac{2\sigma^2}{2} = \sigma^2 \quad \Rightarrow \quad b[P(\nu_k)] = 0, \\ \text{Var}[P(\nu_k)] &= \text{Var}\left[\frac{\sigma^2 \chi_2^2}{2}\right] = \sigma^4 \frac{2 \times 2}{2^2} = \sigma^4. \end{aligned} \tag{2.45}$$

In the case of a WGN regularly sampled, the periodogram is strictly (and not only asymptotically) unbiased. The left panel of Fig. 2.7 verifies (2.45) for $M = 10^4$ Monte Carlo (MC) simulations of a WGN $X(t) \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ of $N = 1024$ points. The empirical values (red and blue points) have been obtained using the estimated mean of the periodogram values at each frequencies (i.e. $\frac{1}{M} \sum_{m=1}^M P_m(\nu_k)$) and the unbiased estimated variance (2.29). We see the good match between the theoretical predictions (superimposed yellow and green lines) and the simulations in both cases with small variations around the theoretical values due to the finite values of used MC simulations.

For the purpose of making a numerical example, we define now an autoregressive process which is a common noise model for colored noise.

Definition 9. Autoregressive process

An autoregressive (AR) model is a correlated process parametrized by an order o and a correlation structure characterized by filter coefficients $\{c_j\}, j = 1, \dots, o$. Such a process, noted $AR(o)$, is generated at a given time t_j according to its previous components and a stochastic term $\epsilon(t_j)$ as:

$$X(t_j) = \sum_{j=1}^o c_j X(t_j - j\Delta t) + \epsilon(t_j) \tag{2.46}$$

with $c_j \in]-1; 1[$ the AR coefficients filter⁵ and $\epsilon(t)$ a white Gaussian series with variance σ^2 . The order and the coefficients are the model's parameters. One representation of this signal is the filtering of a WGN with variance σ^2 by a filter characterised by a transfer function. The theoretical PSD of an $AR(o)$ process is [Brockwell and Davis, 1991]:

$$S_{E,AR}(\nu; \theta_{AR}) := \frac{\sigma^2}{\left| 1 + \sum_{j=1}^o c_j e^{-2\pi i j \nu} \right|^2}, \tag{2.47}$$

where $\theta_{AR} := [o, c_j, \sigma^2]^\top$ is the AR parameter vector. An AR process corresponds to a frequency dependent noise.

Example 8. Periodogram of a colored noise: bias and variance

The right panel of Fig. 2.7 shows the bias and variance of $P(\nu_k)$ with 10^4 Monte Carlo (MC) simulations of an autoregressive (AR) process of order $o = 2$, $\sigma^2 = 1$ and coefficients $\mathbf{c} = [0.8, -0.5]^\top$. The color code is the same as for the case of the WGN (left panel). The PSD of such noise (S_E) is given in (2.47). Even for a relatively small value of N ($= 1024$ here), one can observe the good match between the asymptotical theoretical bias (which is 0) and variance (S_E^2) of the periodogram and the simulations.

In conclusion, when the sampling is regular, the theory of the distribution of P is well understood for periodic signals in colored noise. Moreover, the asymptotic theory seems to be quite accurate even for relatively low value of N . This last point will be further investigated in Chap. 3 (cf. Sec. 3.5.1).

⁵ $c_j \in]-1; 1[$ for the stationary condition of the process.

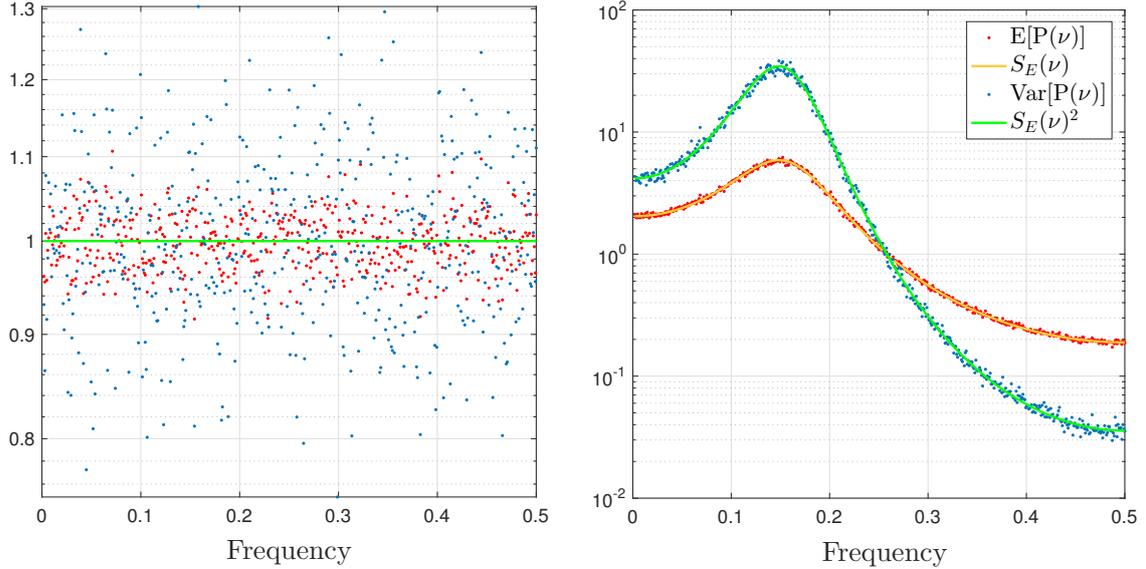


Figure 2.7 – Verification of the asymptotical periodogram bias (yellow solid line) and variance (green solid line) for a WGN $\overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ (left) and for an AR(2) noise (right) with 10^4 MC simulations (red and blue dots). The noise PSD S_E is given by $\sigma^2 = 1$ in the first case and by (2.47) in the second case.

2.2.8 A note on some modified periodograms

A lot of modified periodograms have been developed in the literature to counteract drawbacks of the classical periodogram (as its inconsistency). Examples of such periodograms are given in [Daniell, 1946, Bartlett, 1950, Blackman and Tukey, 1958, Welch, 1967, Capon, 1969, Priestley, 1981, Li, 2014].

Among them, one can distinguish the *averaged periodogram* introduced to reduce the periodogram variance [Bartlett, 1950]. The averaged periodogram consists in averaging L classical periodograms (2.24) to produce a consistent PSD estimate with a reduced variance (factor $1/L$). To built this periodogram, Bartlett has proposed to split the initial time series $\mathbf{X}(t)$ into L different blocks $\{\mathbf{X}_\ell\}$, for $\ell = 1, \dots, L$, assuming the data blocks are uncorrelated and to compute an averaged periodogram as:

$$\bar{P}_L(\nu) := \frac{1}{L} \sum_{\ell=1}^L \frac{1}{N} \left| \sum_{j=1}^N X_\ell(t_j) e^{-i2\pi\nu j \Delta t} \right|^2. \quad (2.48)$$

In our study (cf. Sec. 3.2), we will use a periodogram closely related to (2.48) but conceptually different: L will not be an effective number of blocks taken on the observed time series, but will come from independent noise realizations coming from some simulations.

Many other modified periodograms follow the Bartlett's method. Some of them allow the time series sequences to overlap [Welch, 1967]. Some use specific spectral windows on each of the time series sequence [Blackman and Tukey, 1958, Priestley, 1981, Li, 2014], and some use local averaging of periodogram values [Daniell, 1946] or adapted band-pass filters [Capon, 1969]. The overlap allows to increase the number of averaged sequences to make a compromise between the variance reduction (due to averaging) and a better resolution (which depends on the spectral window of the sequences).

For the case of irregular sampling, the sampled exponentials involved in the FT of the classical and these modified periodograms are not orthogonal. In this case, in the literature of astronomical times series analysis, periodograms adapted to the case of uneven sampling, such as the famous Lomb-Scargle periodogram [Lomb, 1976, Scargle, 1982] have been designed. However, the ordinates of such “adapted periodograms” are dependent (even for a WGN) leading to difficulties to derive the distribution of any test statistics based on their ordinates. This point will be discussed in more details in Chap. 4 (cf. Sec. 4.1 to 4.4).

2.3 Planetary signatures in the periodogram

We study now the radial velocity signatures described in (1.3), first in the time domain and then in the periodogram. In this section, the deterministic part $R(t)$ of model (2.20) is considered. The main lines of this section have been published in [Sulis et al., 2016a].

2.3.1 Signatures in the time domain

The radial velocity of a planet depends on six Keplerian parameters (plus some observational parameters as the inclination, cf. Sec. 1.2.1). Fig. 2.8 illustrates their influence through the independent variations of K_p (through M_p), T_p , e_p , ω_p , t_0 and V_0 for a solar mass star. In each panel, the fixed parameters are $M_p = 1 M_\oplus$, $T_p = 30$ days, $e_p = 0$ (except for the panel c), where $e_p = 0.9$), $\omega_p = 0$ rad, $t_0 = 0$ day and $V_0 = 0$ m.s⁻¹. One can observe the following points:

- The planet mass (panel a)) influences the RV semi-amplitude (in (1.2), $M_p \ll M_\star$ leads to $K_p \propto M_p + \mathcal{O}((M_\star + M_p)^{-2/3})$). The RV amplitude increases with the planet mass ($M_p = [1, 2] M_\oplus$ resp. for the red and black curves). This facilitates the detection of massive planets.
- The planet period (panel b)) influences both K_p ($\propto T_p^{-\frac{1}{3}}$) in (1.2) and the periodicity of the signature ($T_p = 8.3$ and 31.1 days resp. for the red and black curves). This facilitate the detection of short period planets.
- The eccentricity (panel c)) changes the shape of the RV curve. The RV semi-amplitude increases with e_p as $K \propto (1 - e^2)^{-1/2}$ in (1.2), which in principe facilitates the detection ($e_p = [0, 0.5, 0.9]$ resp. for the black, red and blue curves). However, a large eccentricity may in practice not be an advantageous because it requires a thin sampling of the RV curve. For example, a time sampling as the one marked with the green circles will show a flat radial velocity and will be blind to the RV shape variations. If the maximum peak is missed, one can easily be wrong on the orbit eccentricity.
- The argument of periastron (panel d)) impacts the shape of the RV curve when the eccentricity becomes large, as (1.3) moves away from a sinusoid (see also panel c)). This term can decrease the signal amplitude in the RV curve ($\omega_p = \pi/2$, red), introduces asymmetries ($\omega_p = \pi/4$, blue) or inverts the RV amplitudes ($\omega_p = \pi$, black).
- The velocity of the system’s barycentre (panel e)) only introduces a RV shift ($V_0 = [-2, 0]$ m.s⁻¹ resp. for the red and black curves).
- The time of the planet’s passage at periastron (panel f)) only produces a time shift ($t_0 = [0, 50]$ days resp. for the red and black curves).

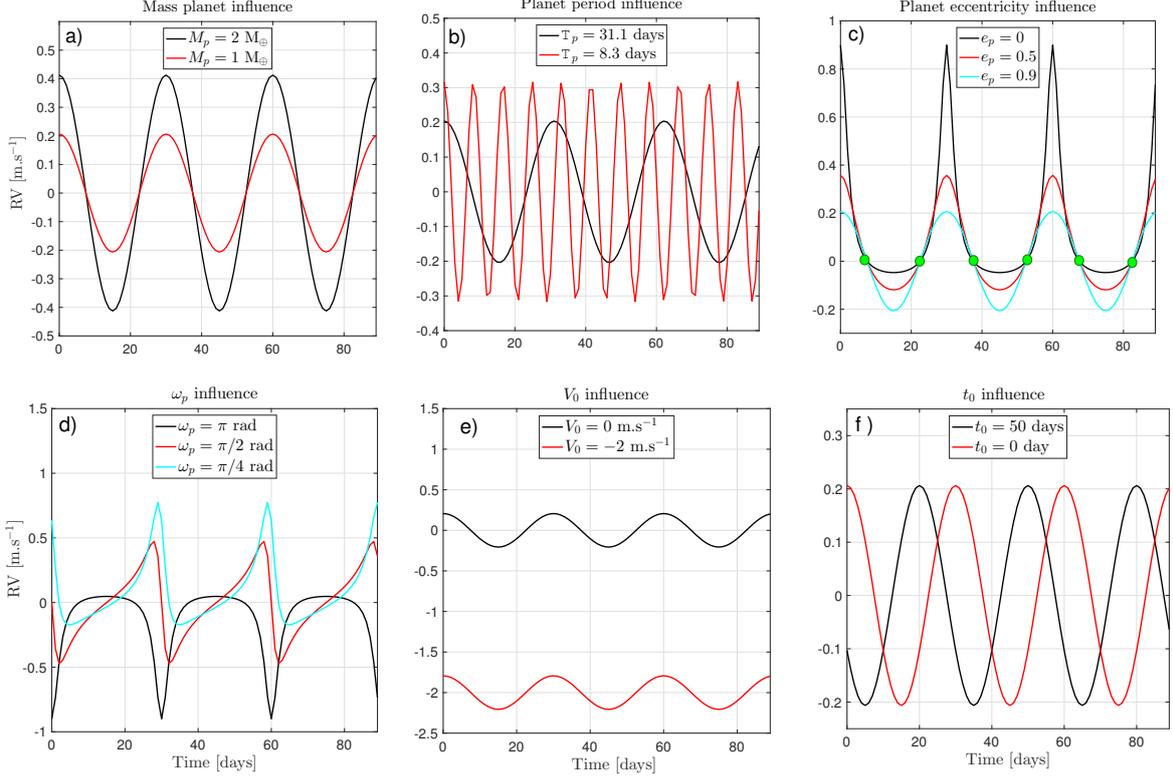


Figure 2.8 – Influence of the Keplerian parameters on RV planet signatures given in (1.3).

2.3.2 Signatures in the Fourier domain

Fig. 2.9 illustrates the influence of the relevant Keplerian parameters one by one in the periodogram. As the velocity of the system barycentre (V_0) and the time of passage at periastron (t_0) do not influence the periodogram at non zero frequencies, we only study here the influence of a) the planet mass, b) period, c) eccentricity and d) the argument of periastron.

In panel a), we overplot on top of the considered “pure” planetary RV signatures a WGN with unit variance: $X(t) = E(t) \stackrel{i.i.d}{\sim} \mathcal{N}(0, 1)$.

This noise could be seen as representative of the instrumental level limit of spectrograph such as HARPS (i.e. 1 m.s^{-1} , cf. Sec. 1.2.2) and is shown in order to compare the planet RV Keplerian amplitudes with the level induced by such random instrumental noise.

The consideration of this “noise level” deserves some attention. In the figure, the black horizontal lines represent three possible ways of measuring this “noise level” in the Fourier domain.

- The average noise level, γ_E , which according to (2.45) is also the variance here:

$$\gamma_E := \mathbb{E}[P] = \sigma_X^2 = 1.$$

This level is illustrated by the dashed line.

- The “90% confidence noise level” of one periodogram value, γ_1 . Using (2.42), the threshold γ_1 is defined, $\forall k$, by

$$\Pr \left[\left(2 \frac{P(\nu_k)}{\sigma^2} \right) < \gamma_1 | \mathcal{H}_0 \right] = \Phi_{\chi^2_2} \left(2 \frac{\gamma_1}{\sigma^2} \right) = \frac{90}{100}.$$

This level is then:

$$\gamma_1 = \frac{\sigma^2}{2} \Phi_{\chi^2_2}^{-1}(0.9),$$

with $\Phi_{\chi_2^2}^{-1}$ the inverse cdf of a chi-square distribution with 2 degrees of freedom. This level is illustrated by the dash-dot line.

- The “global 90% confidence noise level”, γ_η , evaluated for $\eta = \frac{N}{2} - 1$ independent values of the periodogram. The threshold γ_η is defined by

$$\Pr \left[\max_k \left(2 \frac{P(\nu_k)}{\sigma^2} \right) < \gamma_\eta | \mathcal{H}_0 \right] = \prod_{k=1}^{\eta} \Pr \left[\left(2 \frac{P(\nu_k)}{\sigma^2} \right) < \gamma_\eta \right] = \Phi_{\chi_2^2} \left(2 \frac{\gamma_\eta}{\sigma^2} \right)^\eta = \frac{90}{100}$$

which leads to

$$\gamma_\eta = \frac{\sigma^2}{2} \Phi_{\chi_2^2}^{-1}(0.9^{1/\eta}).$$

This level is illustrated by the dotted line.

Note that the periodogram of a pure noise may, by chance, show quite large values: here, $\gamma_E = 1$, $\gamma_1 \approx 2.30$ while $\gamma_\eta \approx 7.43$ for $\eta = 178$ (i.e. $N = 358$ samples).

The parameters of these simulations are $T = 5$ years and $\Delta t = 5.1$ days. While the studied variations of the parameters are indicated by the legend, the fixed parameters are $\omega_p = 0$ rad, $T_p = 364.9$ days (frequency $1/T_p$ on grid, panels a and b) and $T_p = 285.07$ days (frequency off grid, panel a, b and c), $e_p = 0$ (panels a, b) and $e_p = 0.9$ (panels c, d), $t_0 = 2$ days, $V_0 = 0$ m.s⁻¹, $M_p = 1 M_\oplus$, $M_\star = 1 M_\odot$, and $i = 90^\circ$.

Panel a: Influence of the planet mass (M_p). The influence of M_p in (1.3) appears only in the expression of the RV semi-amplitude (1.2). In the figure, different masses are investigated: the blue, red and green curves are associated respectively to $M_p = 2, 1$ and $0.5 M_\oplus$. In this plot, the signal frequency is on the Fourier frequency grid ($T_p = 364.9$ days, corresponding to $\nu_5 = 3.1 \cdot 10^{-8}$ Hz). The grey stems are one periodogram realization of a WGN with unit variance. In comparison with the perturbations caused by an instrument like HARPS, one sees the three planets have signatures of magnitudes comparable to the considered noise levels. The amplitude of the RV signal varies by a factor of 2 and 4 depending on the planet mass. In the Fourier domain, the corresponding periodogram peaks scale as K_p^2 (see (2.25)), so the signal to be detected scales as the squared amplitude. Note that this signature will also increase linearly with N (through D_N (2.26)).

Panel b: Influence of the planet period (T_p). The blue and red curves are associated to $T_p = 364.9$ and 285.07 days. We superimpose the continuous periodograms (in practice the periodogram is oversampled by a factor of 100) associated to these periods (red and blue dashed curves). Two effects play a role here. First, as $K_p \propto T_p^{-\frac{1}{3}}$ in (1.2), the amplitude decreases with the period (K_p is respectively 0.089 and 0.097 m.s⁻¹ for $T_p = 364.9$ and 285.07 days). Second, the planet frequency $f_p = 1/T_p$ may or may not be on the Fourier grid, with the possible power leakage discussed in the previous subsection (see Fig. 2.5). At worst, the amplitude reduction is when f_p falls in the middle of a Fourier frequency interval, as this leads to values not crossing the zeros of the Fejér kernel (cf. (2.25)). In this panel (and the two following ones), the noise levels (horizontal black curves) are shown in order to visually compare the RV amplitudes.

Panel c: Influence of the planet eccentricity (e_p). The blue, red and green curves are associated respectively to $e_p = 0.9, 0.5$ and 0 . To better show the influence of the considered parameter, we consider here an off-grid planet frequency as in panel b. In the case of high eccentricity, the figure shows the apparition of multiple harmonics. This case differs from the situation when a pure sinusoid is present, which corresponds to a circular planetary orbit ($e_p = 0$). However, planets with high-eccentricity orbits are rare and, even for

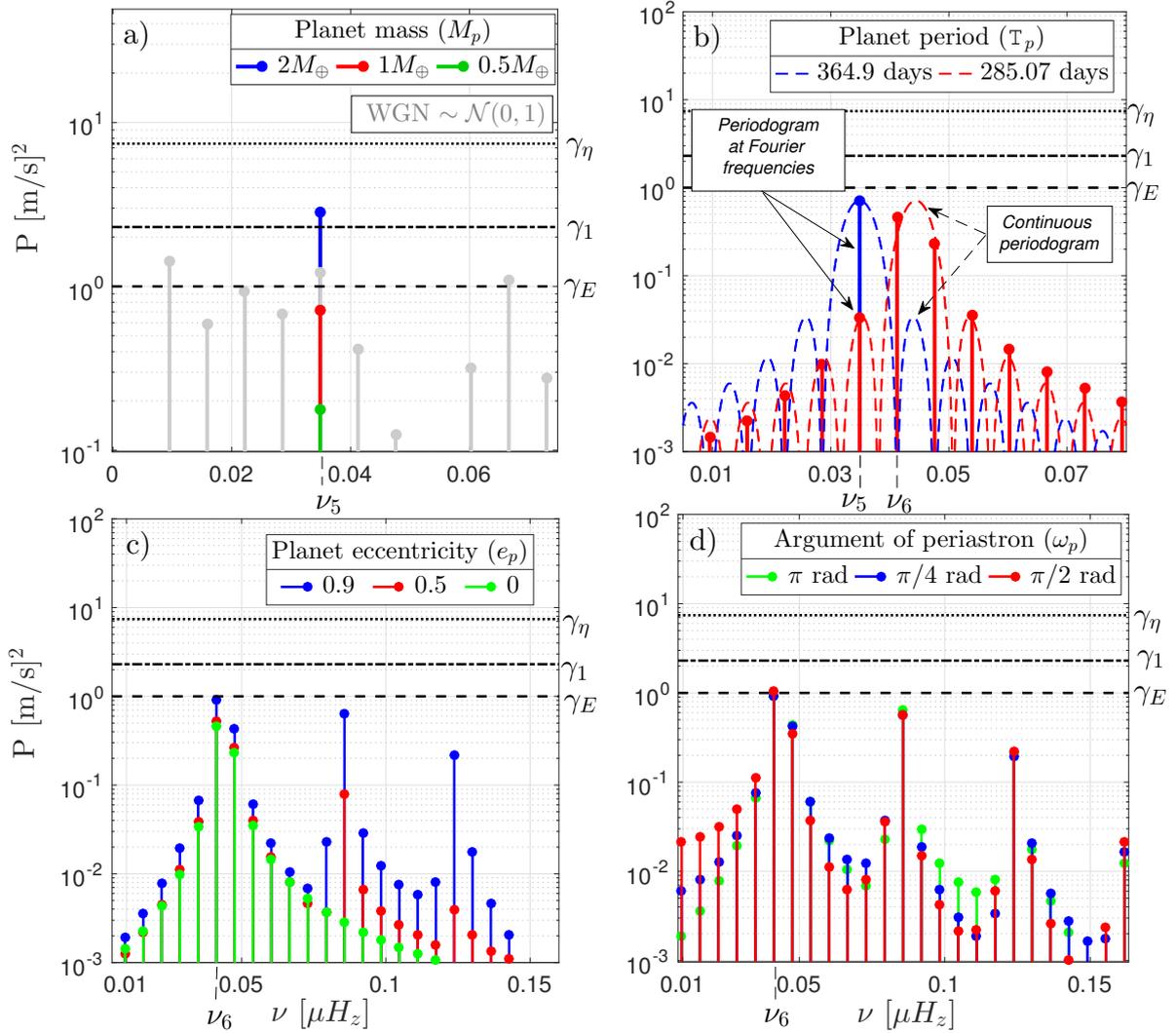


Figure 2.9 – Study of RV planet signatures in the Fourier domain as a function of the planet a) mass, b) period (for a given observational sampling), c) eccentricity and d) the argument of periastron. In panels b, c, and d, one can observe the apparition of harmonics in the periodogram [Sulis et al., 2016a].

the highest eccentricities, the amplitude of the harmonics decay rapidly (the periodogram plot is here in log scale). For example, in the Solar System the highest eccentricity is $e = 0.058$ for Mercury; for exoplanets, the majority of the discovered planets by RV have $e < 0.5$ (cf. right panel of Fig. 2.10).

Panel d: Influence of the planet passage at periastron (ω_p). This term impacts the first term in (1.3). In the last panel, the green, blue and red curves correspond respectively to $\omega_p = \pi, \frac{\pi}{4}$ and $\frac{\pi}{2}$ rad. The considered signal frequency is on the Fourier grid ($T_p = 364.9$ days) and the eccentricity is set to $e_p = 0.9$. One can observe that despite the wild asymmetries of the RV light curves according to the value of ω_p (cf. panel d of Fig. 2.9), their signatures in the periodogram are very similar for all the ω_p considered.

Number of planets (N_p , not shown here). This parameter obviously impacts directly the number of “large” components in the Fourier spectrum through the sum in (1.3). This number can however safely be considered in the range 1 to 10 (say) for the vast majority of exoplanetary systems. One can note that, currently (June 2017), exoplanetary systems found with the largest number of planets have seven planets (these systems are HD 10180 [Olsen and Bohr, 2010], Kepler-90 [Ford et al., 2011], HR 8832 [Vogt et al., 2015], and TRAPPIST-1 [Gillon et al., 2017]). The left panel of Fig. 2.10 shows the number of stars as function of the number of found planets within their System (all detection methods considered). The Solar System is the only one known, at the present time, with eight planets.

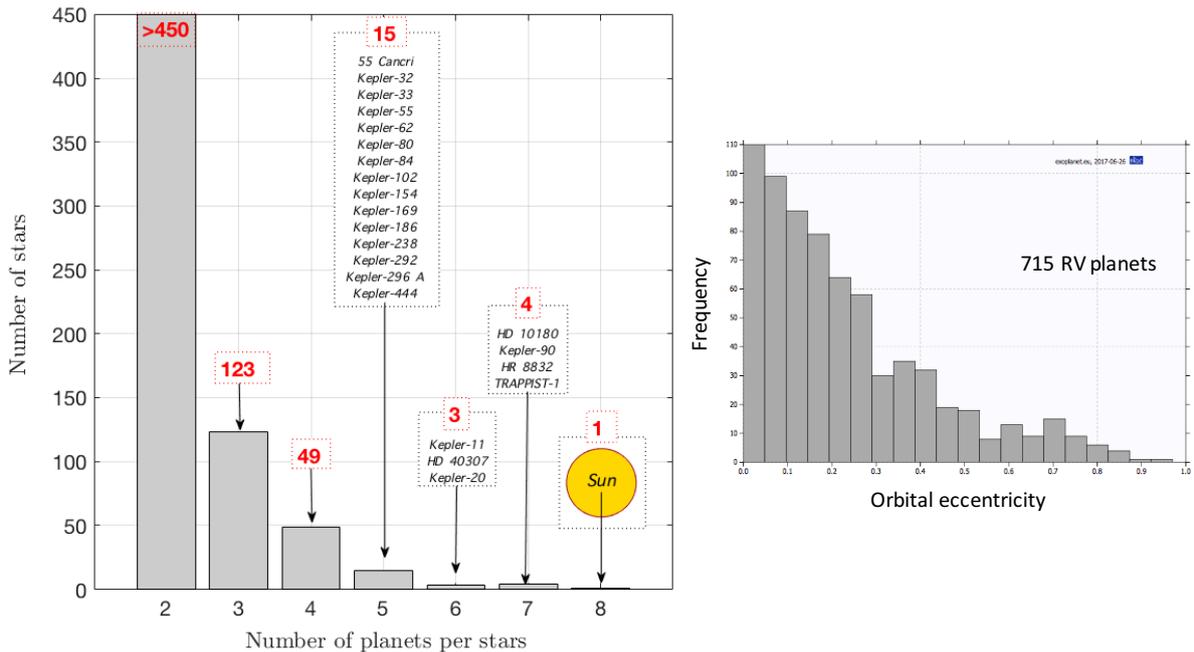


Figure 2.10 – Left: Histogram of the discovered multi-planetary system as function of the number of planets (all detection methods considered). Right: Histogram of the planets orbital eccentricities (only RV detection). Source: www.exoplanets.eu, June 2017

2.3.3 Discussion and signal model

To summarise, we have seen that planets with RV of a few tens cm.s^{-1} reach amplitudes in periodogram at levels comparable to the sample noise level discussed in Sec. 2.3.2. The detection of the planet signature appears indeed difficult unless very long time series are available. The peaks amplitudes can further be decreased for off-grid planet frequencies and the spectral leakage effect has always to be taken into account. Numerous peaks may arise in the periodogram because of off-grid frequencies and/or high eccentricities.

Strictly speaking, the number of nonzero Fourier coefficients for the periodogram of a pure RV signature always equals the total number of Fourier frequencies, N , even for a pure sine (zero eccentricity) as there is a null probability for the signal frequency to fall exactly on the Fourier grid. According to (2.25), the periodogram amplitude should be close to $\frac{NK^2}{2}$ in this case. However, we have seen that in all cases (except perhaps very rare and exotic systems), the RV spectral signatures exhibit only a small fraction of significant harmonics with respect to the noise level (this is because the planets around a host star are in small number and tend to have eccentricities well below 1). In short, while the RV spectrum of low mass planets can have complicated structures, it is sparse⁶ in the vast majority of cases (though not strictly sparse). Consequently, RV signals can be modelled by a sum of an unknown (but small) number of pure sinusoids. In other words, we make the approximation that the Keplerian RV planet signatures can be described by a finite sum of coefficients of a Fourier series.

In the following, we call this number N_S , and we say that $N_S \ll N$. When there is one planet with frequency close to the Fourier grid and low eccentricity, N_S will be essentially 1. This leads to the model hypotheses:

$$\begin{cases} \mathcal{H}_0 : X(t_j) = E(t_j) \\ \mathcal{H}_1 : X(t_j) = \sum_{q=1}^{N_S} \alpha_q \sin(2\pi f_q t_j + \varphi_q) + E(t_j) \end{cases} \quad (2.49)$$

with $E(t_j)$ the considered noise. Under \mathcal{H}_1 , $X(t_j)$ contains an unknown (but small) number of sinusoidal components (N_S) with unknown parameters $\boldsymbol{\theta}_q := [\alpha_q, f_q, \varphi_q]^\top$. Note that the α_q may not be equal to the RV semi-amplitudes K described in (1.2) (this is only the case when (1.3) becomes a sine function, i.e. when the planetary orbit is circular). Note also that if several planets are present, N_S encapsulates all the corresponding harmonic signals.

Now, what is the relevance of a detection strategy based on approximate model (2.49) instead of exact RV signatures ?

Assuming $R(t_j) = V_R(t_j)$ in (1.3), an alternative approach following closely the Keplerian RV model would be to discretise the Keplerian parameter space and end-up with a large library of Keplerian profiles, which could be tested in a matched filter approach. Because such a dictionary would necessarily be redundant (as all the Keplerian parameters have to be considered), the resulting “periodogram” ordinates would not be independent and controlling the false alarm would be delicate. Similarly, *Keplerian periodograms* (e.g. [Cumming, 2004]) consist in a least-square fit of function (1.3) to the observed data (cf. Chap. 4, Sec. 4.2).

In contrast, the discrete Fourier transform (through the periodogram) acts as an orthonormal analysis dictionary of elementary harmonic oscillations, with well studied properties allowing for analytical approximations (at least asymptotically as seen in Sec. 2.2 in (2.34), (2.35) and (2.39)) and, importantly, fast algorithms. Note also that from a detection point of view, it may

⁶By definition, a signal is said *sparse* (resp. strictly sparse) when most of its coefficients are approximatively (resp. strictly) null in a given *dictionary*. A dictionary is a representation of the observed signal. In our case, it is the orthogonal basis of the discrete exponentials of the DFT.

not be more powerful to use a large library of signals representing accurately the possible alternative signals than testing for only a reduced number (even a single) of approximate signals. The reason is that by increasing the number of atoms of the library, the false alarm rate is also increased. An example of this effect is described in the first pages of [Suleiman et al., 2014].

A classical approach in detection could be to use the GLR test defined in (2.19) with a Keplerian RV model. However, deriving a GLR test by maximizing, as in (2.19), the likelihood with the exact RV model over all unknown parameters seems out of reach because

- (i) the complicated interdependence of some parameters (e.g. terms M_p , e_p , T_p in (1.2) and (1.3)) would make the maximization difficult, even numerically;
- (ii) the unknown number of planets, even small, would require to ally the GLR to model selection procedures, which in turn would make it difficult to evaluate the resulting false alarm rate.

These considerations suggest that opting for a simple multi-sinusoidal model and combining the (classical) periodogram ordinates for testing purposes allows to set up tests of good power with the possibility of evaluating analytically their performance even when the noise PSD S_E is partially unknown (see Chap. 3) with, in addition, very fast detection algorithms. In Chap. 3, we will show that for extrasolar detections by RV, the test of the maximum periodogram component, described in (2.51) for $N_S = 1$ should be a good choice in many situations but adaptive procedures allowing to deal with sparse (but not extremely sparse) signals ($1 < N_S \ll N$) should also be considered for more complex Keplerian signatures (like HC or BJ, cf. (2.70) and (2.74)).

We finally note that even if a full GLR approach will not be considered among the tests used in this thesis, some of them are GLR tests in some particular cases of (2.49) (e.g. the Fisher's test, defined in (2.58) is a GLR when the noise is white Gaussian under the null and $N_S = 1$ under the alternative).

2.4 Statistical tests applicable to the periodogram

This section aims to present some tests that can be applied to the periodogram ordinates. They have been selected for reference as they cover different model settings (i.e. particular cases of model (2.49)). Many of these tests have been designed in their original form for the simplest case of a white noise under the null and one contribution of this thesis is to extend their application to the case of partially unknown colored noise. Under some assumptions on the distribution of the considered variates, the properties of these tests applied to the periodogram are well known and are summarized below.

2.4.1 Preliminary notations

These tests are better presented using order statistics, which for a vector $\mathbf{Z} = [Z_1, Z_2, \dots, Z_N]^\top$ will be denoted by:

$$\min_k Z_k := Z_{(1)} < Z_{(2)} < \dots < Z_{(N)} := \max_k Z_k.$$

In the sequel, \mathbf{Z} will indeed be taken as the vector of ordinates of periodogram \mathbf{P} (2.24) or as a normalized version of \mathbf{P} (see Sec. 2.6). We present the tests on a generic data vector \mathbf{Z} as the considered tests can be applied to any data vector. Moreover, the application of some of them to the periodogram was not considered yet in the literature.

All tests below are of the form $T(\mathbf{Z}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma$, with $T(\mathbf{Z})$ the test statistic and $\gamma \in \mathbb{R}^+$ a threshold that determines the false alarm rate.

Also, recall from (2.12) that the observed P -value v_T corresponding to the realization $t := T(x)$ of the random variable $T(X)$ is defined as:

$$v_T(t) := \Pr(T(X) \geq t | \mathcal{H}_0).$$

2.4.2 Classical tests

Schuster's test

The first test that has been applied to the periodogram and that is the ancestor of all the other tests presented here, is the Schuster's test [Schuster, 1898]

$$T_S(\mathbf{Z}_k) := \mathbf{Z}(k) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \quad (2.50)$$

where k corresponds to a predefined component of \mathbf{Z} . This test has been developed for a model involving under \mathcal{H}_0 a white Gaussian noise of known variance σ^2 and under \mathcal{H}_1 a single complex signal with unknown amplitude but known frequency $f_s = \nu_k$ falling on the Fourier grid.

Considering $\mathbf{Z} := 2\mathbf{P}/\sigma^2$, the ordinates of the periodogram evaluated at successive Fourier frequencies in Ω are (even non asymptotically) i.i.d. with known distributions under both hypotheses and given respectively in (2.42) and (2.43). In this case, the P_{FA} can be easily written as

$$\begin{aligned} P_{FA}(T_S(2\mathbf{P}(\nu_k)/\sigma^2), \gamma) &:= \Pr(2\mathbf{P}(\nu_k)/\sigma^2 > \gamma | \mathcal{H}_0) \\ &= 1 - \Phi_{\chi_2^2}(\gamma) \\ &= 1 - (1 - e^{-\gamma}) \\ &= e^{-\gamma}. \end{aligned}$$

Schuster's test (2.50) is a binary hypothesis test in which a single component of \mathbf{P} is tested. As in practice the frequency of interest is unknown (and generally not falling on the Fourier grid), other test statistics have been devised, as the test of the maximum.

Test of the maximum

This test statistics is based on the maximum component:

$$T_M(\mathbf{Z}) := Z_{(N)}. \quad (2.51)$$

For N independent variates Z_i of cdf Φ_{Z_i} , using (2.4) one can deduce the false alarm of the test $T_M(\mathbf{Z}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma$:

$$P_{FA}(T_M(\mathbf{Z}), \gamma) = 1 - \Pr[\text{"All the components } Z_i \text{ are below } \gamma"] = 1 - \prod_{i=1}^N \Phi_{Z_i}(\gamma).$$

When \mathbf{Z} is the periodogram, a particular set of frequencies $\Omega' \subset \Omega$ is often considered. The alternative becomes "there is at least one sinusoidal signal with a frequency in Ω' ".

For a model (2.20) involving under \mathcal{H}_0 a white Gaussian noise of known variance σ^2 , the probability of false alarm writes, for $\Omega' = \Omega$:

$$\begin{aligned} P_{FA}(T_M(2\mathbf{P}/\sigma^2), \gamma) &:= \Pr(T_M(2\mathbf{P}/\sigma^2) > \gamma | \mathcal{H}_0) \\ &= 1 - \prod_{k \in \Omega} \Pr\left(\frac{2P_k}{\sigma^2} \leq \gamma | \mathcal{H}_0\right) \\ &= 1 - \left(\Phi_{\chi_2^2}(\gamma)\right)^{\frac{N}{2}-1} \\ &= 1 - (1 - e^{-\gamma})^{\frac{N}{2}-1}, \end{aligned} \quad (2.52)$$

with $\Phi_{\chi_2^2}$ the cdf of a χ_2^2 r.v. and Ω the indices set defined in Sec. 2.1.

For a model (2.20) involving under \mathcal{H}_1 a complex signal as in (2.38) with unknown frequency (but on the Fourier grid) and under \mathcal{H}_0 a WGN of known variance, using (2.3), the probability of detection writes:

$$P_{DET}(T_M(2\mathbf{P}/\sigma^2), \gamma) := \Pr(T_M(2\mathbf{P}/\sigma^2) > \gamma | \mathcal{H}_1) \approx 1 - \prod_{k \in \Omega} \Phi_{\chi_{2, \lambda_k}^2}(\gamma) \quad (2.53)$$

with λ_k the non centrality parameter defined in (2.40). In this case, test $T_M(\mathbf{P})$ corresponds to the GLR test described in (2.19) [Kay, 1998].

Test of the variance

An interesting test to compare with T_M is the test of the variance. Considering $\mathbf{Z} = \mathbf{P}$, this test is interesting because the variance of the time series is proportional to the sum of the periodogram components by the Parseval's identity (2.33).

For a model (2.20) involving under \mathcal{H}_0 a zero mean white Gaussian noise of known variance σ^2 , using (2.42), this test statistics writes

$$T_{VAR}(\mathbf{P}) := \frac{2}{N-1} \sum_{k=1}^{N/2-1} P_k^2, \quad (2.54)$$

and is proportional to an estimate of the time series variances. The distribution of this test statistic is

$$T_{VAR}(\mathbf{P}) \sim \frac{\sigma^2}{2} \frac{2}{N-1} \sum_{k=1}^{N/2-1} \chi_2^2. \quad (2.55)$$

Consequently, this test is related to $\frac{N}{2} - 1$ independent random variables χ_2^2 distributed. The sum of these r.v. follows a χ^2 distribution with $2 \times (\frac{N}{2} - 1) = N - 2$ degrees of freedom leading to

$$T_{VAR}(\mathbf{P}) \sim \frac{\sigma^2}{N-1} \chi_{N-2}^2.$$

The P_{FA} of this test can then be deduced as

$$\begin{aligned} P_{FA}(\gamma, T_{VAR}(\mathbf{P})) &= \Pr(T_{VAR}(\mathbf{P}) > \gamma | \mathcal{H}_0) \\ &= 1 - \Pr(T_{VAR}(\mathbf{P}) \leq \gamma | \mathcal{H}_0) \\ &= 1 - \Phi_{\chi_{N-2}^2} \left(\gamma \frac{N-1}{\sigma^2} \right). \end{aligned} \quad (2.56)$$

Under \mathcal{H}_1 , using (2.39) for a model involving a complex signal as in (2.38), the P_{DET} writes

$$\begin{aligned} P_{DET}(\gamma, T_{VAR}(\mathbf{P})) &= \Pr(T_{VAR}(\mathbf{P}) > \gamma | \mathcal{H}_1) \\ &= 1 - \Pr(T_{VAR}(\mathbf{P}) \leq \gamma | \mathcal{H}_1) \\ &= 1 - \Phi_{\chi_{N-2, \lambda_k}^2} \left(\gamma \frac{N-1}{\sigma^2} \right), \end{aligned} \quad (2.57)$$

with $\gamma(P_{FA}) = \frac{\sigma^2}{N-1} \Phi_{\chi_{N-2, \lambda_k}^2}^{-1}(1 - P_{FA})$.

Example 9. WGN case: T_M vs T_{VAR}

Consider model (2.20) with under \mathcal{H}_0 a WGN $\overset{i.i.d}{\sim} \mathcal{N}(0, 1)$ of $N = 1024$ data points and under \mathcal{H}_1 a complex signal of the form (2.38) in such WGN. The signal frequency is located on the Fourier grid ($f_s = \nu_{667} = 0.15$) and the amplitude is $\alpha_s = 0.15$.

Fig. 2.11 compares the tests $T_M(\mathbf{P}/\sigma^2)$ (black) given in (2.51) and $T_{VAR}(\mathbf{P}/\sigma^2)$ (red) given in (2.54). In the panels a) and b), the theoretical $P_{FA}(\gamma)$ for the both tests (solid lines) given in (2.52) and (2.57) are compared with their empirical estimates (dotted lines) obtained through 10^5 MC simulations. One can observe the good match validating the theoretical expressions. In panel c), the ROC curves of both tests, computed using equations (2.52), (2.53), (2.56) and (2.57), are shown for the considered model. One can observe that T_M outperforms T_{VAR} for all P_{FA} . The last panel represents $P_{DET}(N; P_{FA} = 0.1)$ for both tests (computed using the relation $\gamma(P_{FA})$ in (2.53) and (2.57)). While these tests show similar behavior for large N , T_M shows better performance for finite N value. This is because the energy in the Fourier domain is mainly concentrated on the signal frequency, leading to a sparse signature easier to detect than when testing on the global dispersion of temporal time series as T_{VAR} .

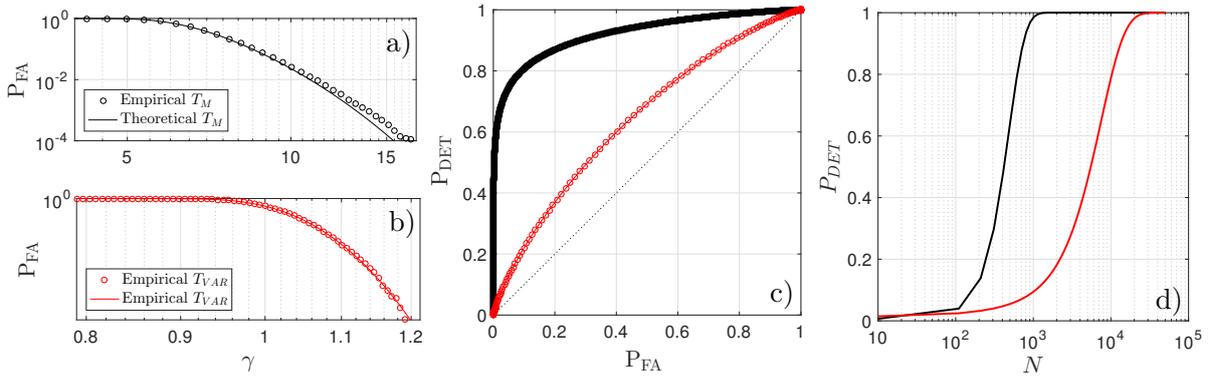


Figure 2.11 – Comparison of the performance of T_M and T_{VAR} for a WGN under \mathcal{H}_0 and a single complex signal under \mathcal{H}_1 . a,b) Verification of the false alarm probabilities expressions given in (2.52) and (2.57) with MC simulations for resp. T_M and T_{VAR} . c) Associated ROC curves. d) Detection probabilities given in (2.53) and (2.57) as function of N .

In the previous expressions of P_{FA} , one notes that we have to know σ^2 to be able to set the detection threshold γ . In practice however, the noise variance / power spectral density is often unknown. Hence, other tests like the Fisher's test have been developed (for WGN) to deal with an estimate of the noise variance σ^2 .

Fisher's test

In 1929, Fisher developed a test for the case where the noise is a WGN of unknown variance [Fisher, 1929]. The test statistic is

$$T_F(\mathbf{Z}) := \frac{Z_{(N)}}{\sum_{k=1}^N Z_{(k)}}. \quad (2.58)$$

If $\mathbf{Z} = \mathbf{P}$, this test consists in normalizing the maximum periodogram component by an asymptotically unbiased estimate of the PSD (up to a known constant that can be absorbed in the

threshold) (cf. (2.33)). Considering frequencies in Ω , Fisher's test applied to $\mathbf{Z} = \mathbf{P}$ writes:

$$T_F(\mathbf{P}) := \frac{P_{(\frac{N}{2}-1)}}{\sum_{k=1}^{\frac{N}{2}-1} P_{(k)}}. \quad (2.59)$$

When a single monochromatic signal, with frequency on grid and unknown parameters $\boldsymbol{\theta}_s := [\alpha_s, f_s, \phi_s]^\top$, is involved \mathcal{H}_1 in a WGN of unknown noise variance, the Fisher's test is the most powerful test [Anderson, 1971]. In such a case, this test corresponds to the GLR test (see [Quinn, 1986], who also covers the case of more than one sinusoid).

Moreover as investigated below for a WGN with unknown variance, the Fisher's test is a Constant False Alarm Rate (CFAR) detector.

Definition 10. Constant False Alarm Rate detector

If a test is able to guarantee a prescribed false alarm rate under the null hypothesis although some parameters under \mathcal{H}_0 (like the noise variance) are not known by the detector, the test is called a Constant False Alarm Rate detector.

The false alarm rate of T_F in case of a WGN is [Fisher, 1929]

$$P_{FA}(T_F(\mathbf{P}), \gamma) = \sum_{j=1}^M (-1)^{j-1} \binom{N/2}{j} \left[1 - j\gamma\right]_+^{\frac{N}{2}-1}, \quad (2.60)$$

where M the largest integer satisfying $1/\gamma < M \leq \frac{N}{2}$ and $\binom{N/2}{j} := \frac{(N/2)!}{j!(N/2-j)!}$ the binomial coefficients. One can clearly note that the P_{FA} of T_F in (2.60) does not depend on σ^2 , which makes the test CFAR.

Examples of works using this test in Astronomy are [Koen, 2015a, Koen, 2015b, Schwarzenberg-Czerny, 1998, Aittokallio et al., 2001, Gutiérrez-Soto et al., 2009].

However, one can note two important drawbacks to this test:

1. The performance of this test decrease for an increasing number of periodic components in the data [Chiu, 1989].
2. When the noise is not a WGN but is colored, the calibration by the sum of the periodogram components does not give a consistent estimate of the noise PSD. Consequently the P_{FA} given in (2.60) is not valid.

The first drawback leads to turn to tests designed for multiple sinusoids as the ones described in the following sections. The second drawback will be discussed in Sec. 2.6.

2.4.3 Tests designed for multiple sinusoids

In the case of multiple sinusoids involved in model (2.49) under \mathcal{H}_1 ($N_S > 1$), the performance of the Fisher's test are known to decrease, owing to the perturbations of sinusoids in the noise variance estimation. Many alternatives tests exist (e.g. [Shimshoni, 1971, Siegel, 1980, Bhansali, 1979, Priestley, 1981, Bölviken, 1983b, Bölviken, 1983a, Chiu, 1989, Truong-Van, 1990, Von Sachs, 1993, Von Sachs, 1994, Artis et al., 2004, Zheng, 2012, Li, 2014]) offering globally the same robustness against an unknown noise variance [McSweeney, 2006].

One approach to deal with multi-sinusoids is to treat them as outliers in the periodogram to better estimate the noise level by excluding a number $N_C \geq 1$ of components presumably contaminated by the sinusoids. This is interesting when one suspects that the time series may contain several pure sinusoids under the alternative. In our application framework, we have seen that this is the case for

1. one single sinusoid with frequency off-grid,
2. N_S pure sinusoids on grid,
3. a planet signature with exotic Keplerian parameters (i.e. high eccentricity or off-grid frequency, cf. Sec. 2.3),
4. N_p signatures different from “pure” sinusoidal signals.

Regarding the fact that testing the largest peak may not be the best strategy for multiple sinusoids, Chiu & Shimshoni have demonstrated that order statistics, different from the maximum, may be more discriminative than $Z_{(N)}$ against the null [Shimshoni, 1971, Chiu, 1989]. For a detailed comparison of tests using *a priori* knowledge on the number of sinusoids, one can refer to [Artis et al., 2004].

T_C test

In this spirit, we propose a test, called T_C below. Assuming a good estimate of N_S is available (say N_C), this test looks at the N_C^{th} order statistics of \mathbf{Z} instead of the highest one:

$$T_C(\mathbf{Z}) := Z_{(N-N_C+1)}. \quad (2.61)$$

The test T_M in (2.51) is a particular case of T_C when $N_C = 1$. Of course, the performance of such a test depends on parameter N_C (in practice, the best performance are not necessarily achieved for $N_C = N_S$; this will be shown in Chap. 3, Sec. 3.6).

For a model (2.20) involving under \mathcal{H}_0 a WGN of known variance, the P_{FA} of T_C can be derived as follows. When $\mathbf{Z} = 2\mathbf{P}/\sigma^2$, consider the case where $N_C = 3$ and $\Omega' = \Omega$. The P_{FA} is

$$P_{FA}(T_C(2\mathbf{P}/\sigma^2)) = 1 - \Pr[\text{“zero } P(\nu_k) \text{ ordinates } > \gamma\text{”}] \\ \dots - \Pr[\text{“exactly one } P(\nu_k) > \gamma\text{”}] - \Pr[\text{“exactly two } P(\nu_k) > \gamma\text{”}].$$

Using the independence of the periodogram ordinates, this leads to

$$\Pr[\text{“exactly two } P(\nu_k) > \gamma\text{”}] = \binom{N/2-1}{2} \times \Pr[\text{“exactly 2 given } P(\nu_k) \text{ ordinates } > \gamma\text{”} | \mathcal{H}_0] \\ \dots \times \Pr[\text{“all the other ordinates } \nu_{k'} \neq \nu_k \text{ are } < \gamma\text{”} | \mathcal{H}_0],$$

and using $\Pr[2P(\nu_k)\sigma^2 > \gamma] = 1 - \Phi_{\chi_2^2}(\gamma)$ in (2.42), we obtain

$$\Pr[\text{“2 } P(\nu_k) > \gamma\text{”}] = \binom{N/2-1}{2} \left[1 - \Phi_{\chi_2^2}(\gamma) \right]^2 \Phi_{\chi_2^2}(\gamma)^{N/2-1-2}.$$

Consequently, for a general value of N_C , the P_{FA} of T_C generalizes to

$$P_{FA}(T_C(2\mathbf{P}/\sigma^2; N_C)) = 1 - \sum_{j=0}^{N_C-1} \binom{N/2-1}{j} \left[1 - \Phi_{\chi_2^2}(\gamma) \right]^j \Phi_{\chi_2^2}(\gamma)^{N/2-1-j}. \quad (2.62)$$

Equation (2.62) reduces to (2.52) for $N_C = 1$.

Chiu's test

In the case of multiple sinusoids in a WGN with unknown variance involved in model (2.20), one can improve over T_C in (2.61) by normalising T_C by the sum of the periodogram components (i.e. by an estimate of the noise variance (2.33)). Chiu proposed to produce a better estimate of the noise variance σ^2 than T_F , in (2.58), by trimming a few of the largest \mathbf{Z} components when computing the average of the remaining ordinates [Chiu, 1989]. The Chiu's test statistics writes

$$T_{Ch}(\mathbf{Z}) := \frac{Z_{(N-N_C+1)}}{\sum_{k=1}^{N-N_C} Z_{(k)}}. \quad (2.63)$$

When the noise is white Gaussian, the P_{FA} of T_{Ch} is the same than for the Fisher's test given in (2.60) [Chiu, 1989].

Robust Fisher test

Following the same idea, the robust Fisher test statistics is defined as (cf. [Li, 2014], Eq.6.2.18)

$$T_{F,rob}(\mathbf{Z}) := b_r N r \frac{Z_{(N)}}{\sum_{k=1}^{N-r} Z_{(k)}}, \quad (2.64)$$

with $r = \frac{N-N_C}{N}$ the fraction of remaining values, and $b_r = 1 + \frac{1-r}{r} \log(1-r)$ a constant ensuring a consistent variance estimate in absence of sinusoids.

In the case of a WGN of unknown variance, the exact distribution of this test statistics is derived in [Bölviken, 1983a]. For $r = 1$, this test is reduced to the Fisher's test (up to a constant N).

Approaches (2.61), (2.63) and (2.64) pose the question of the choice of N_C , which must be set *a priori*. These tests have decreasing power in case of strong mismatch between the value of parameter N_C . Not fixing N_C in advance but estimating this parameter from the data may lead to more powerful tests, but at the cost of a more difficult control of the FA rate (as N_C becomes random). This suggests to consider other approaches that are adaptive in the number of sinusoids. This is the case of the two tests below: the Higher Criticism and the Berk-Jones tests.

Example 10. Comparison of the considered tests

Consider model (2.20) with a WGN $E(t) \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ of unknown variance σ^2 under the null, and a sinusoidal signal of the form $R(t) = \sum_{i=1}^{N_S} 0.5 \sin(2\pi f_i t)$ under the alternative with f_i the N_S signal frequencies randomly chosen in the Fourier grid. The time series length is $N = 100$. The purpose of this example is to demonstrate the robustness of tests T_F (2.58), T_{Ch} (2.63) and $T_{F,rob}$ (2.64) by studying: a) their performance for a WGN of unknown variance depending on the number of sinusoids, b) the variation of the multi-sines test performance with regard to the setting of parameter N_C and c) the behavior of such performance in a low order AR colored noise.

Panel a) of Fig. 2.12 illustrates the empirical ROC curves of the Fisher (green), Chiu (red) and robust Fisher tests (blue) obtained with 10^4 MC realizations of $X(t)$ for a single sinusoid ($N_S = 1$) and for multiple sinusoids ($N_S = 10$) under \mathcal{H}_1 . The tests are equivalent for the case $N_S = 1$ (shown in black). For the case $N_S = 10$, the Chiu and robust Fisher tests are evaluated for $N_C = N_S$. One can observe that i) T_F is not the best, ii) compared to T_F , $T_{F,rob}$ is the best at large P_{FA} , and iii) at small P_{FA} both T_{Ch} and $T_{F,rob}$ outperform T_F .

Panel b) illustrates the ROC curves of the Chiu and Fisher tests for different N_C values and $N_S = 10$. The considered values are $N_C = [5, N_S, 20]$ (resp. the dashed, solid and dotted curves). One can observe variations in the performance of these tests according to the value of N_C parameter. The $T_{F,rob}$ performance increase with the three considered values of N_C , while T_{Ch} shows opposite performance. Moreover, the $T_{F,rob}$ performance variate slowly with the N_C value in contrast to the ones of T_{Ch} . While the T_{Ch} performance vary substantially, $T_{F,rob}$ shows relatively stable performance whatever N_C . On this example, the $T_{F,rob}$ performance are less influenced by the choice of the setting parameter N_C .

Panel c) shows the non-reliability of all these tests when the noise is colored (here illustrated by the AR(2) process used in Example. 8). All the ROC are located under the ROC's diagonal meaning that the tests are meaningless in this case. The ROC are completely influenced by the noise PSD. The FA (and detection) probability of such tests are not valid (e.g. (2.60)) and so the tests cannot be reliably implemented in this case.

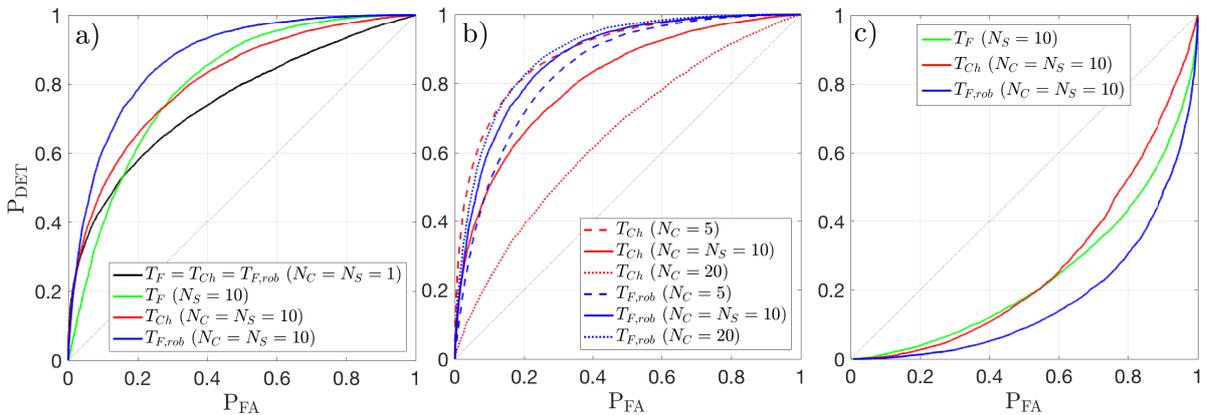


Figure 2.12 – ROC curves of tests T_F (2.58), T_{Ch} (2.63) and $T_{F,rob}$ (2.64) for $N_S = 10$ sinusoids in noise. a) Tests comparison for a WGN and multi-sines tests parametrized with $N_C = N_S$. b) Influence of N_C on multi-sines tests performance. c) Tests comparison for a colored noise (AR(2)).

2.5 Adaptive tests that can be applied to the periodogram

Some adaptive tests have recently been designed in order to detect faint sparse vectors means in i.i.d. normal vectors. They are designed for the so-called “rare and weak detection” problem [Donoho and Jin, 2004, Ingster et al., 2010, Walther, 2012, Arias-Castro et al., 2011]. To our knowledge, such tests have not been applied yet to the periodogram.

These adaptive tests can be classified in the family of goodness-of-fit (GOF) tests. GOF tests can be divided in two groups:

1. Tests based on the first moments of \mathbf{Z} with respect to an orthonormal basis. This group is let out of scope of this study but examples of such tests are the Neyman’s smooth test [Neyman, 1937], the Ledwina’s test [Ledwina, 1994], or the Rayner’s test [Rayner et al., 2010].
2. Tests based on measures of deviations between the theoretical and empirical \mathbf{Z} distribution under \mathcal{H}_0 . Each measure of mismatch corresponds to a test. Examples of such tests are: Cramer-von Mises (and its generalization, the Anderson-Darling test) [Anderson and Darling, 1954], Kolmogorov-Smirnov [Kolmogorov, 1933, Smirnov, 1948, Berk and Jones, 1979], Berk-Jones [Berk and Jones, 1979], Phi-divergence [Jager and Wellner, 2007], and the Higher Criticism [Tukey, 1976].

In this manuscript, we only focus on this second family for which a brief description is proposed below.

2.5.1 Introduction of the problem

Adaptive tests can be introduced by considering a problem involving under the null a process \mathbf{Z} for which the ordinates $\{Z_k\}$ are i.i.d. with known marginal distribution (considered standard Gaussian below) and under the alternative a sparse mean vector

$$\begin{cases} \mathcal{H}_0 : Z_k \sim \mathcal{N}(0, 1), \\ \mathcal{H}_1 : Z_k \sim \mathcal{N}(\mu_k, 1), \quad \mu_k > 0, \text{ for some } k \in [1, \dots, N]. \end{cases} \quad (2.65)$$

Often, the sparsity is measured under \mathcal{H}_1 by the proportion of nonzero coefficients as (cf. [Donoho and Jin, 2004])

$$\epsilon := \frac{\#\{\mu_k \neq 0\}}{N} := N^{-\beta}, \quad \text{for } \beta \in]0, 1]. \quad (2.66)$$

Since the ordinates Z_k are independent, the P -values are i.i.d and uniformly distributed. For a vector of r.v. $\mathbf{Z} = [Z_1, Z_2, \dots, Z_N]^\top$ of which $\mathbf{z} = [z_1, z_2, \dots, z_N]^\top$ is one realization, the P -values (2.12) are thus of the form:

$$v_k = \Pr(\mathcal{N}(0, 1) > z_k) = 1 - \Phi(z_k),$$

with Φ the standard gaussian cdf.

2.5.2 Example: Kolmogorov-Smirnov test

As before, denote

$$\min_k z_k := z_{(1)} < z_{(2)} < \dots < z_{(N)} := \max_k z_k$$

the ordered values of \mathbf{z} and by $Z_{(1)}, \dots, Z_{(N)}$ the order statistics of \mathbf{Z} .

The observed P -values corresponding to \mathbf{z} will be denoted by $v_{\mathbf{z},k}$ (with $v_{\mathbf{z},k} := \Pr(Z_k > z_k)$) and the observed ordered P -values by $v_{\mathbf{z},(k)}$. The corresponding r.v. will be denoted by $V_{\mathbf{z},k}$

and $V_{\mathbf{Z},(k)}$.

GOF tests are based on the important property that the P -values are independent and their distribution under the null is uniform. Let us denote by Φ_V the cdf of a uniform r.v. ($\Phi_V(\gamma) = \gamma$). The corresponding empirical cdf of the P -values is given by:

$$\widehat{\Phi}_V(\gamma) := \frac{\#\{v_{\mathbf{Z},k} \leq \gamma\}}{N}. \quad (2.67)$$

The idea of the considered GOF test is to measure a deviation between Φ_V and $\widehat{\Phi}_V$ and to reject the null if the difference is large.

As an example, the Kolmogorov-Smirnov test statistics measures the maximum absolute deviation between the observed and expected P -values cdf [Berk and Jones, 1979]

$$KS(\mathbf{Z}) := \sup_{\gamma} |\widehat{\Phi}_V(\gamma) - \gamma|. \quad (2.68)$$

Based on the KS test statistics (2.68), Anderson & Darling introduced a weight function to calibrate the difference in (2.68) [Anderson and Darling, 1954]. Some years later, the *Higher Criticism* (HC) was introduced in the same spirit [Tukey, 1976].

2.5.3 Higher Criticism

Also called second-level testing, the Higher Criticism test compares the fraction of observed test statistics above a threshold to the expected fraction under the null hypothesis [Blomberg, 2012]. Donoho & Jin defined the test HC, which first computes weighted deviations of the empirical cdf of the P -values w.r.t. their theoretical (uniform) cdf under the null [Donoho and Jin, 2004]:

$$HC_k := \frac{\sqrt{N}(k/N - v_{\mathbf{Z},(k)})}{\sqrt{v_{\mathbf{Z},(k)}(1 - v_{\mathbf{Z},(k)})}}. \quad (2.69)$$

The test statistic is the maximum of those weighted deviations (2.69):

$$HC(\mathbf{Z}) := \max_{1 \leq k \leq \alpha_0 N} HC_k \quad (2.70)$$

with $\alpha_0 \in [\frac{1}{N}, 1]$ a parameter.

The functioning of HC is illustrated by the left panel of Fig. 2.13 for $N = 5$ and $\alpha_0 = 1$ (adapted from [Paris, 2013]). The green line represents the true cdf of the P -values and the blue discontinuous curve illustrates an example of an empirical cdf. The red arrows indicates the scores HC_k in (2.69) (up to the denominator factor of (2.69)). For example here, the highest component of Z corresponds to the P -value $v_{(5)}$, which has the largest score (HC_5).

The parameter α_0 has to be set according to the data degree of sparsity under the considered alternative hypothesis. This parameter expresses the fact that computing all ratios in (2.70) is not useful since the considered alternative is sparse. In principle, α_0 can be set in the interval $[1/N, 1/2]$, corresponding to the interval containing the smallest P -values (i.e. the largest \mathbf{Z} components) and the median P -value. Generally, authors choose $\alpha_0 = 1/2$ without a significant impact on the test behaviour.

The theoretical (asymptotic) analysis of the detection problem posed by (2.65) was made by [Ingster et al., 1998, Jin, 2004]. Following their notations, they parametrize the model (2.65) under \mathcal{H}_1 by

$$\begin{cases} \epsilon := N^{-\beta}, & \text{for } 1/2 < \beta < 1, \\ \mu := \sqrt{2r \log(N)}, & \text{for } 0 < r < 1, \end{cases} \quad (2.71)$$

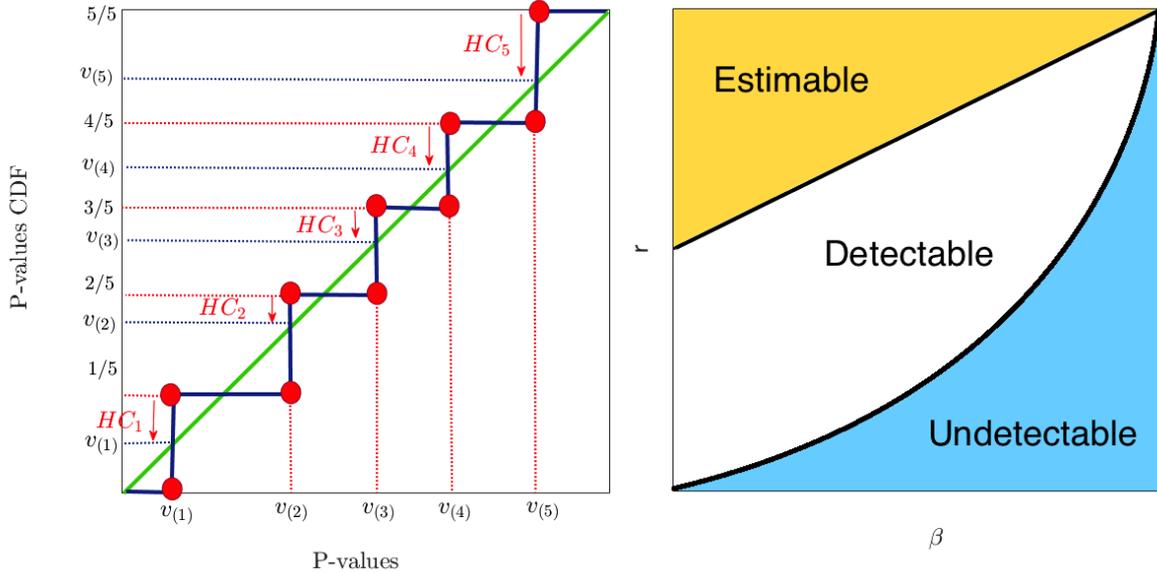


Figure 2.13 – Left: Illustration of the principle of the HC test statistics defined in (2.70). Right: Illustration of the detectability region defined in (2.72).

with $\beta = 1$ modelling a single sparse signal and $\beta = 1/2$ modelling a non extremely sparse signal. These authors found a *threshold curve*, also called a *detection boundary curve*, in the plane of the sparsity domain (through β) and the signal(s) amplitude r . This function separates the detectable regions from the undetectable ones and writes

$$\rho^*(\beta) = \begin{cases} 0, & \text{for } 0 < \beta < 1/2, \\ \beta - 1/2, & \text{for } 1/2 \geq \beta < 3/4, \\ (1 - \sqrt{1 - \beta})^2, & \text{for } 3/4 \geq \beta < 1, \end{cases} \quad (2.72)$$

A blueprint illustrating the principle of a detection boundary curve is shown in the right panel of Fig. 2.13 (extracted from [Donoho and Jin, 2004]). The detection boundary separates the detectable to undetectable regions. In the estimable region, it is possible to consistently estimate the components. In some words, their results show that:

- In the detectable region ($r > \rho^*(\beta)$), one can adjust the Neyman-Pearson test to achieve an asymptotic null probability of error (i.e. $P_M + P_{FA} \rightarrow 0$).
- In the undetectable region ($r < \rho^*(\beta)$), it is not possible to separate the testing model hypotheses and the sum of error tends to 1.

They proved that for $r > \rho^*(\beta)$, the HC statistics has full power asymptotically. This is interesting as HC does not need knowledge of the sparsity parameter ϵ nor the signal parameter μ . According to Donoho & Jin, HC has better asymptotic power characteristics than the max test T_M , defined in (2.51), for a specific range of contamination number (assuming they have same amplitude).

Finally, as the HC_k in (2.69) are dependent, one cannot compute analytically the false alarm of this test for finite N but this is possible numerically. For example, efficient algorithms for computing significance levels (and hence the function $\gamma \mapsto P_{FA}(\gamma)$) of HC for finite (but possibly large) values of N can be found in [Moscovich et al., 2016, Moscovich-Eiger and Nadler, 2015].

Let us now see how HC fits in our detection problem. For a model (2.20) involving a WGN $\overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ under the null with known variance σ^2 , the vector $\mathbf{Z} = 2\mathbf{P}/\sigma^2$ has i.i.d. ordinates distributed as in (2.42). The ordered P -values involved in (2.69) are thus ordered values of P -values computed by

$$v_{2P/\sigma^2, k} := \begin{cases} 1 - \Phi_{\chi_2^2}(2 P(\nu_k)/\sigma^2), & \forall k \in \Omega, \\ 1 - \Phi_{\chi_1^2}(2 P(\nu_k)/\sigma^2), & \text{for } k = 0, \frac{N}{2}. \end{cases} \quad (2.73)$$

Under the alternative (exoplanetary signatures), a fraction of the ordinates contain a deterministic part, and hence follows the non central χ_2^2 distribution given in (2.39). The frequencies at which these deviations occur are unknown. Their magnitudes (related to the non centrality λ_k) are also unknown and weak, in the sense that they are comparable to the expected magnitude of the periodogram maximum under the null hypothesis. So this is a rare and weak detection problem, for which HC is indeed very relevant.

Example 11. HC applied to $\mathbf{Z} = 2\mathbf{P}/\sigma^2$

Consider model (2.20) with a WGN $E(t) \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ under the null and $N_S = 40$ sines under the alternative. All the sines have a unit amplitude and the frequencies have been taken randomly in the set of Fourier frequencies. The purpose of this example is to illustrate the HC test statistics given in (2.69) when applied to $2\mathbf{P}/\sigma^2$, with the P -values computed as in (2.73).

Fig. 2.14 illustrates the ordered statistics $Z_{(k)}$ (left), the ordered P -values $v_{Z,(k)}$ (middle) and statistics HC_k (right) as a function of k/N . The time series has 500 points, leading to 249 independent periodogram components following the χ_2^2 distribution (2.39). The parameter $\alpha_0 = 1$ here. The red vertical line indicates the max value of HC_k (right panel) found in $k = 41$ (i.e. $k/N = 0.165$) with $HC_{41} = 2.7$. So here the test statistics would be HC_{41} and \mathcal{H}_0 would be rejected if HC_{41} is greater than the test threshold. Note that in this case the test naturally identifies a number of 41 large periodogram ordinates. In this example, one could say that there are 41 deviations under the alternative, which are the 41 highest periodogram peaks. This corresponds here to all $Z_{(k)} \geq 4.398$ (left) or all $v_{Z,(k)} < 0.11$ (middle). In this example, this number is close to the true N_S .

To summarize, optimality⁷ results of HC are asymptotic and established for a specific sparsity *vs* amplitude model. When the deviation occurs at a single frequency or for extremely sparse signatures, Donoho & Jin showed that the test based on the maximum value (T_M , given in (2.51)) is asymptotically optimal in the sense that whenever the Neyman-Pearson test has full power, the T_M test has full power as well. In such cases, the largest \mathbf{Z} ordinate (or, equivalently, the smallest P -value) is the most powerful test statistic to discriminate against the null hypothesis. For sparse, but not extremely sparse signatures, adaptive tests are more powerful as not focused only on the highest periodogram components (as also discussed about the classical tests T_{Ch} , $T_{F,rob}$ and T_C before) [Donoho and Jin, 2004]. Regarding the problem of detecting RV planet signatures or their harmonics, the HC test can be very interesting. We will come back to this point in Chap.3 (cf. Sec. 3.5.2).

⁷A test is said optimal in [Donoho and Jin, 2004], if the sum of the probability of error and the probability of missed detection tends to zero for large N .

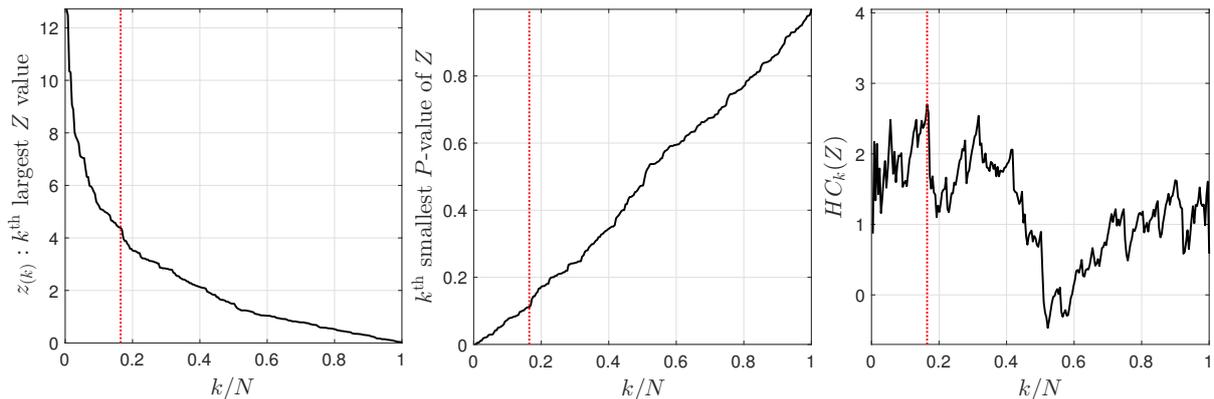


Figure 2.14 – Illustration of HC test statistics (2.70) applied to $N_S = 40$ sines in a WGN through a) the ordered periodogram peaks, b) ordered P -values $v_{Z,(k)}$ and c) HC_k statistics defined in (2.69) as a function of k/N (i.e. the expected cdf of the P -values). The red vertical lines indicate the k/N component corresponding to the max of HC_k , i.e. $k = 41$.

2.5.4 Berk-Jones test

For HC (2.70), the deviations in (2.69) are established using the asymptotic convergence of a binomial distribution to a Gaussian distribution. In the tails, however, this convergence is very slow (see e.g., [Mary and Ferrari, 2014, Li and Siegmund, 2015, Moscovich et al., 2016] for illustrations). For this reason the test statistics below was recently (and almost simultaneously) proposed by [Mary and Ferrari, 2014, Aldor-Noiman et al., 2013, Kaplan and Goldman, 2014, Gontscharuk et al., 2014]. As noted in [Moscovich et al., 2016, Gontscharuk et al., 2016], this test was initially proposed by Berk and Jones (and called M_n^+) in [Berk and Jones, 1979].

This test is based on exact significance reflected by the P -values, that is, on the P -values of the ordered P -values. Since the ordered P -values are ordered statistics from a uniform distribution, they are beta distributed [David and Nagaraja, 2003], with

$$V_{\mathbf{Z}} \sim B(k, N - k + 1),$$

their P -values involve the cdf of beta variables, which is a regularized incomplete beta function given in (2.9) :

$$\Pr(B(k, N - k + 1) \leq x) = I_x(k, N - k + 1).$$

This leads to a test statistic, related to HC, called the Berk-Jones (BJ) test statistics [Moscovich et al., 2016, Aldor-Noiman et al., 2013, Mary and Ferrari, 2014, Kaplan and Goldman, 2014, Gontscharuk et al., 2014]:

$$BJ(\mathbf{Z}) := \max_{1 \leq k \leq \alpha_0 N} I_{1-v_{\mathbf{Z},(k)}}(N - k + 1, k). \quad (2.74)$$

As for HC, one can apply this test to the periodogram. In this case, the P -values are computed as in (2.73).

BJ presents the same adaptive optimality as HC for sparse mixture detection, and the asymptotic distribution of the BJ test statistic can be found in the Theorem 4.1 of [Moscovich et al., 2016]. As for HC, convergence to the asymptotic distribution may be slow. Efficient algorithms for computing significance levels of BJ can be found (as previously noted for HC) in [Moscovich et al., 2016, Moscovich-Eiger and Nadler, 2015].

An important point is the applicability of such adaptive tests in case of colored noise under \mathcal{H}_0 . When the noise is colored with autocovariance function Σ^{-1} , some attempts exist, assuming

Σ^{-1} is known, to compute the adaptive tests (e.g. [Hall and Jin, 2010]).

To conclude, adaptive tests such as HC and BJ are indeed interesting for periodicity detection. However, they are not in nominal regime when the noise is colored with unknown statistics as the periodogram P -values can not be easily computed. We will see in Chap. 3 that it is still possible to build implementable adaptive (and CFAR) detectors in this case. Although adaptive tests like HC and BJ have been mostly evaluated and designed for large N values, we will also show in Chap. 3 (cf. Sec. 3.5.2) that they can be very powerful for finite N values as well. One can finally note that, while the utility of such adaptive tests has been evoked (and will be considered in Chap. 3) in the case of detection by RV data, they could also be very interesting for detecting planetary signatures by the transit method. Indeed, detecting planetary transits is also a “rare and weak” detection problem, where multiple periodic deviations (stellar light flux diminution) occur under the alternative. In the case where P -values can be computed, i.e. when we have access to the noise statistics, the application of adaptive tests as HC (2.70) and (2.74) could be made directly on the time series for analysis. As my thesis attempts to control these statistics through independent noise simulations of the colored noise of interest, the application of adaptive tests to photometric data is one of the main perspectives of this work.

2.6 Sinusoid detection in unknown noise statistics: related works

This section aims to discuss the problem of controlling the distribution of tests based on the periodogram ordinates when the noise is colored with unknown statistics.

So far, we have mostly illustrated the detection tests behavior for time series of WGN (with known or unknown variance). If the noise E in model (2.49) is colored with known PSD $S_E(\nu_k) \neq \sigma^2$, all the performance calculations for P_{FA} and P_{DET} developed for a WGN for T_M (2.51), T_C (2.61), T_F (2.58), T_{Ch} (2.63) and $T_{F,rob}$ (2.64) are still (asymptotically) valid by replacing the ordinates $Z_k = P(\nu_k)/\sigma^2$ by $Z_k = P(\nu_k)/S_E(\nu_k)$ with $P(\nu_k)/S_E(\nu_k) \sim \chi^2_2/2$ under the null $\forall k \in \Omega$ according to (2.35).

When the noise has unknown statistics, the distribution of P under \mathcal{H}_0 given in (2.35) is not known and consequently the significance level (the size) of the test (as those described in Sec. 2.1 to 2.4) is not known either. In this case, different approaches can be followed:

1. The first approach is simply to ignore possible noise correlations and to apply tests designed for white noise. However, as was illustrated in Fig. 2.12 (right panel), the statistical behavior of the resulting testing procedure may be hazardous, with unpredictable significance level and poor power (see also [Kay, 1999] and [Sulis et al., 2016b] on this point).
2. The second approach consists in cleaning the data by a whitening procedure. This procedure consists to find the dominant peak within the periodogram, fit a sine using the found frequency and subtract it to the data. This is repeated until the dominant peak of the periodogram residuals is found at the noise level. The resulting time series is then considered as “whitened”. In astronomy, this method is largely applied (under the name of *pre-whitening* [Hatzes, 2013]) to remove the periodogram frequencies presumably affected by (stellar) noise. Obviously, this method is always hazardous when the noise statistics are unknown. Characterising precisely the statistical distribution of the resulting cleaned residual components is consequently often a problem.

3. The last approach consists of estimating the noise PSD so that the hypothesis of a known noise statistics is considered to hold approximately. This estimate of the noise PSD can be used to calibrate the periodogram of the data $P(\nu)$, leading to a frequency-wise standardized periodogram of the form

$$\tilde{P}(\nu|\hat{S}_E) := \frac{P(\nu)}{\hat{S}_E(\nu)}. \quad (2.75)$$

Such kind of standardization can be found, for example, in [Lu and Zimmerman, 2005, Li, 2014]. The standardization by \hat{S}_E (i.e., a r.v., instead of S_E) changes the distribution of $P(\nu_k)/S_E(\nu_k)$, which is different from a χ_2^2 distribution. Standardization (2.75) can be seen as a generalization of the Fisher's test approach (2.58) (which standardizes the periodogram ordinates by the sample mean of the periodogram ordinates, i.e. an estimate of the PSD of WGN).

The estimate $\hat{S}_E(\nu)$ can be

- Non parametric, i.e. the noise PSD is estimated without assuming any particular model on the PSD structure. However, other parameters may be involved in the estimation procedure as, for example, parameters of specific spectral windows [Hannan, 1961, Priestley, 1981]
- Parametric, i.e. the noise PSD depends (in a known manner) on a finite number of parameters (e.g. AR models).

These methods estimate the model's parameters from the observed data, which can lead to three main problems for controlling the distribution of the test statistics:

- (a) the choice of the noise PSD model,
- (b) the estimation errors related to the parameters estimation of the chosen model,
- (c) the possible bias on the parameters' estimation due to the presence of the (unknown) signal of interest.

Errors on the noise modelization or the parameters estimation can reduce the power of the test and produce an over (or under) estimation of \hat{S}_E , which can lead to false (or missed) detections.

2.6.1 Non parametric approaches

Whittle suggests a method to deal with the problem of detecting sinusoids in mixed spectra (i.e., oscillating signal plus random noise) [Whittle, 1954]. He proposed a truncated periodogram to estimate the noise PSD as in (2.75):

$$\hat{S}_E(\nu_k) = \frac{1}{2\pi} \sum_{s=-(m-1)}^{m-1} \hat{r}_E(s) e^{-i\nu_k s} \quad (2.76)$$

with m a truncation point parameter and \hat{r}_E the noise auto-covariance function estimated on selected data components using AR models. Using (2.76), he designed a statistic test based on the Fisher test (2.58), which is:

$$T_W(\mathbf{P}) := \frac{\max_k P_{(k)}/2\pi\hat{S}_E(\nu_k)}{\sum_{k=1}^{N/2} P_{(k)}/2\pi\hat{S}_E(\nu_k)}. \quad (2.77)$$

The test statistics T_W in (2.77) follows the same distribution as T_F (2.58) with a P_{FA} given in (2.60) [Whittle, 1954]. In practice, this test is known to present two “weaknesses”. First, its performance suffer from the possible presence of several sinusoidal signals in the data, as they influence the estimation of $\hat{S}_E(\nu_k)$ and thus the power of T_W . Second, the test omits some pre-selected data components in evaluating \hat{r}_E (through parameter m in (2.76)) assuming they are due to harmonics *before* the establishment of the existence of such harmonics. This is a strange assumption when the purpose of the test is precisely to determine whether harmonics are present or not.

Bartlett proposed to use “grouped periodograms” to estimate $\hat{S}_E(\nu_k)$ assuming an approximately constant PSD over some frequency bands of given length [Bartlett, 1955]. These grouped periodograms are parametrized by the segment length setting, which could deteriorate the test performance in case of strong noise correlations [Priestley, 1981].

Following the reasoning of Whittle’s test T_W (without using an AR noise however), Hannan proposed to use a spectral window on the periodogram to attenuate the influence of ordinates possibly contaminated by sinusoids [Hannan, 1961]. Hannan defined a “truncated periodogram window” (W_m) as a weighted mean of the periodogram components:

$$\hat{S}_E(\nu_k, \boldsymbol{\theta}) = \frac{1}{N} \sum_{t=-N+1}^{N-1} \left(1 - \frac{|t|}{N}\right) r_E(t) W_m(\nu_k - \boldsymbol{\theta}) e^{-i2\pi\nu t}, \quad (2.78)$$

with $W_m(\theta)$ a spectral window depending on parameters $\boldsymbol{\theta}$ and a “truncation point” m . The window has to be chosen to remove the possible signal contributions at a given Fourier ν_k by smoothing the periodogram over the frequency bins. Hannan proposed several spectral windows according to different assumptions on the data. Moreover, he showed that the resulting estimate \hat{S}_E is an unbiased estimate of the noise PSD even if a signal influences the tested frequencies. However, some problems of estimate (2.78) appear due to the dependence on the spectral leakage effect and of the large variability observed in practice in presence of multiple harmonics under \mathcal{H}_1 [Nicholls, 1967, Nicholls, 1969].

An other method is the Priestley’s “double window periodogram” based on the autocovariance of the data (Chap. 8 of [Priestley, 1981], [Bhansali, 1979]). This method is also called the *correlogram* approach. Priestley assumes the existence of an integer $m \ll N$ such that the auto-covariance of the noise $r_E \approx 0$ for $|n| \geq m$ and evaluates:

$$P_{w_n, w_m}(\nu_k) = \frac{1}{2\pi} \sum_{s=-N+1}^{N-1} \left(w_n(s, \nu_k) - w_m(s, \nu_k)\right) \hat{r}_E(\nu_k) \quad (2.79)$$

with w_n and w_m two different lag windows, n and m their respective truncation points and \hat{r}_E the covariance function estimated from the data. After the evaluation of (2.79), Priestley’s method divides the range of the positive frequencies into smaller intervals, for which the cumulative sum of (2.79) is computed. Then, a test is defined as the ratio of this cumulated sum in (2.79) standardized by a function of the used spectral windows w_n and w_m . The resulting test is shown to be normally distributed under the condition that $\hat{r}_E \rightarrow 0$ for large N . Knowing the distribution of this statistical test, Priestley proposed an iterative method based on the residuals of (2.79) (i.e. in which the largest component of the Priestley’s test statistics has been removed) to estimate the number of harmonics \hat{N}_S within the data. The iterations are stopped when the statistical test, normally distributed under the null, is inferior to a given threshold (iteratively decreased by a factor $1/\hat{N}_S^{(i)}$ with $\hat{N}_S^{(i)}$ the counter of found harmonic components in the Priestley’s iterative procedure).

While the distribution under \mathcal{H}_0 of the standardized PSD estimate in (2.75) derived by these methods can be accessed assuming $\hat{S}_E \approx S_E$ (in such case the P_{FA} is given by (2.60)), the reliability of this approximation cannot really be checked in practice. The reader is referred to Priestley's book for more details about all these tests (Chap.8 of [Priestley, 1981]).

Other methods that are adaptive for the case of colored noise exist. Such methods use regression techniques with robust M-estimators [Von Sachs, 1993, Von Sachs, 1994] or still model selection procedures based on information theoretic criteria [Kavalieris and E.J., 1994]. Techniques reducing the influence of signal peaks under the alternative are proposed in [Chiu, 1990] using peak-insensitive Yule-Walker spectral estimates. [Kay, 1999] proposes a method combining a smoothed periodogram PSD estimate and the GLR test (2.19), under the assumption of a known broadband signal in unknown colored noise. Different approaches, related to standardizations of the kind of (2.75) also exist: ratios of pairs of periodogram ordinates for symmetry testing purposes [Lu and Zimmerman, 2005], of periodogram ordinates over distinct frequency bands for detecting a signal with specified frequency [Liavas et al., 1998], the use a geometric mean (instead of arithmetic mean in Fisher's test) for standardization [Zheng, 2012].

As a general rule, when following GLR approaches for detecting multiple sinusoids in unknown number, the GLR must be combined with model selection procedures. While sharp model selection criteria and CFAR detectors exist under white noise, the correlated case remains an open problem [Nadler and Kontorovich, 2011].

2.6.2 Parametric methods

Parametric methods consist in assuming specific models for the noise PSD. Under this assumption, their goal is to robustly estimate the parameters of this model. Common models are stationary autoregressive or ARMA (AR Moving Average) processes. Such methods are iterative procedures based on AR model of periodogram residuals [Priestley, 1981, Quinn and Fernandes, 1991, Quinn, 1999], ARMA models [Gryca, 1998, Truong-Van, 1990] or still methods using ratios of AR process as spectral estimates [White, 1999]. Many methods exist to estimate the parameters of such AR or ARMA models. For instance, when using an estimate based on an AR process of order o , this order can be estimated using many criteria, *e.g.* [Akaike, 1969, Akaike, 1974, Parzen, 1975, Hannan and Quinn, 1979, Rissanen, 1984, Ding et al., 2015]. We will be more specific on some of these models in the next chapter (cf. Sec. 3.4).

Even if such approaches are relatively straightforward to implement, characterising the distribution of $\tilde{P}(\nu_k|\hat{S}_E)$ in (2.75) is very difficult owing essentially to the stochastic nature of the estimated model parameters. The difficulty in this case is the dependence of the distribution of \hat{S}_E on estimated parameters, which complicates the analytical characterization of the distribution of $\tilde{P}|\hat{S}_E$. Moreover, as for the case of non parametric method, further complications arise when multiple sinusoids are present under the alternative, as they perturb the estimation of model of the noise PSD parameters. For colored noise with unknown PSD, we are not aware of works studying the false alarm rate when AR/ARMA or other models are used for test standardization as in (2.75).

2.7 Conclusions

2.7.1 Main conclusions

Under some conditions which are

- (i) The sampling is regular;
- (ii) The periodogram is evaluated at Fourier frequencies (ν_k in (2.22));
- (iii) The number of samples N is large (asymptotic regime);
- (iv) Under \mathcal{H}_0 , the random process X is stationary Gaussian with known statistics.

the periodogram is a powerful tool, with well understood statistical properties (at least asymptotically).

Our study of the RV planetary signatures in the periodogram (Sec. 2.3) has shown that, for a large majority of Keplerian parameters, these signatures are sparse in the frequency domain (with exceptions for large orbital eccentricities). In practice, the number of harmonics that are significant with respect to the noise level is small because the planets are in small number and tend to eccentricities well below 1. In short, the RV detection problem can be casted as a multiple sinusoids detection problem, for which a few number of sinusoidal variations are present under \mathcal{H}_1 . In the time domain, a relevant model is thus (2.49):

$$\begin{cases} \mathcal{H}_0 : X(t_j) = E(t_j) \\ \mathcal{H}_1 : X(t_j) = \sum_{q=1}^{N_S} \alpha_q \sin(2\pi f_q t_j + \varphi_q) + E(t_j) \end{cases} \quad (2.80)$$

This chapter also provided an extended (though unavoidably selective) overview of classical and recent techniques in sinusoid detection, with emphasis on the problem of designing CFAR detectors for the composite hypotheses of multiple sinusoids and colored noise.

Depending on our knowledge of the noise statistics, if the noise is white or colored, we have considered different statistical tests that can be applied to the periodogram ordinates (cf. Table. 2.5).

Adaptive tests are particularly interesting for the field of extrasolar planets detection consisting of detecting non extremely sparse signatures in a large number of frequencies. To our knowledge, such approaches are for the moment unknown to the astronomy community, which mainly employs the classical test T_M based on the highest periodogram ordinates.

As we have seen, at least for regularly sampled observations, the significance level of all tests applied to the periodogram can be obtained when the noise statistics is known for colored noise (e.g. T_M , T_C , cf. column 7 of Table. 2.5) and for white noise (e.g. T_F , T_{Ch} , $T_{F,rob}$, cf. column 5). When the noise has unknown statistics, two cases can be considered:

- For white noise of unknown variance, several studies provide accurate results for standardized test statistics of the form (2.75), *e.g.*, [Fisher, 1929, Shimshoni, 1971, Siegel, 1980, Bölviken, 1983b, Bölviken, 1983a, Chiu, 1989] (cf. column 6).
- For colored noise with unknown PSD, we are not aware of works studying the false alarm rate when parametric models are used for test standardization. When non parametric models are used, procedures described in [Hannan, 1961, Bhansali, 1979, Priestley, 1981] provide asymptotic control of the false alarm rate. These procedures do not operate explicitly on test statistics of the form (2.75) but rather on windowed periodograms that

TESTS		$R(t)$ in model (2.80)		$E(t)$ in model (2.80)			
Name	Equation	N_S	Free parameter	WGN		Colored Noise	
				σ^2 known	σ^2 unknown	S_E known	S_E unknown
T_S	(2.50)	$N_S = 1$	None	Yes	No	Yes	No
T_M	(2.51)	$N_S = 1$	None	Yes	No	Yes	No
T_{VAR}	(2.54)	$N_S = N$	None	Yes	No	Yes	No
T_F	(2.58)	$N_S = 1$	None	Yes	Yes	No	No
T_C	(2.61)	$N_S \geq 1$	N_C	Yes	No	Yes	No
T_{Ch}	(2.63)	$N_S \geq 1$	N_C	Yes	Yes	No	No
$T_{F,rob}$	(2.64)						
KS	(2.68)						
HC	(2.70)	$1 \leq N_S \ll N$	α_0	Yes	No	Yes	No
BJ	(2.74)						
1	2	3	4	5	6	7	8

Table 2.5 – Table of the some test statistics found in the literature for white noise (Sec. 2.4). In the last four columns, answers “Yes/No” indicate if the P_{FA} control is possible for the considered noise model under the null.

depend on several parameters. We found that these tests are in practice sensitive to parameter setting and that estimating these parameters impact the significance level at which the tests are conducted. This level can of course be approximated by simply neglecting the influence of such a “preprocessing stage” (dealing, *e.g.*, with model order selection, filtering, adaptive window design, or standardization). For instance, we might pretend that $\hat{S}_E = S_E$ as in (2.75) [Whittle, 1952] (which is at least asymptotically true if \hat{S}_E is a consistent estimate). As it will be highlighted in Chap. 3 (cf. Sec. 3.5.4), the actual significance level obtained when doing so may however be far from the assumed one. This leaves open the question of designing both powerful and CFAR tests for unknown colored noise.

In the particular field of exoplanet detection using radial velocity data, one can find similar families of techniques as the ones cited in this chapter. Assuming a White Gaussian noise case, several works use standardized test in the wake of Fisher’s approach (e.g. [Koen, 2015a, Koen, 2015b]). To account for unknown colored noise, tests for planet detection are in practice done after “whitening” the noise through preprocessing stages (e.g., local trend filtering or harmonic removal), which exploit an *a priori* knowledge about known or presumed noise sources (e.g., instrumental drifts or phenomena driven by stellar rotation, like spots) [Hatzes, 2013]. Parametric noise models are also used, as ARMA processes (e.g. [Tuomi et al., 2012]) or more specific empirical models of stellar power spectral density based for example on the Harvey’s laws given in (1.6) [Harvey, 1985, Dumusque et al., 2011].

2.7.2 This thesis

As explained in Chap. 1, a major difficulty for detecting exoplanet signatures in RV data is the stellar noise. Noise is always adversarial, but the situation is much worse when the noise statistics are not known. Power losses (missed detections) or in contrary false detections may occur at rates that are completely uncontrollable for the astronomer when looking for RV targets.

The approach followed in this thesis is thus to seek to take advantage from the realism of recent 3D stellar hydrodynamical simulations (described in Sec. 1.4). These complex

simulations account for numerous processes at the stellar surface: hydrodynamics, radiative transfer, ionisation, dissociation of molecules and magnetic field. They are *ab initio*, i.e. they only depend on the stellar characteristics (T_{eff} , $\log g$, $[\text{Fe}/\text{H}]$), which can be measured by several independent methods (e.g. asteroseismology, spectroscopy). Specifically, the idea is to exploit such simulations of the stellar activity to “calibrate” the periodogram of the RV data, similarly as in (2.75), in order to control the statistical significance of the testing procedure.

One main objective of this thesis is to design a detection method that does not rely on any noise estimation from the data (exploiting stellar MHD simulations detailed in Sec. 1.4), cleaning procedure or parametric noise models with parameters to set.

The simple but relevant model considered under \mathcal{H}_1 and the theoretical properties of the simple but powerful analysis tool constituted by the periodogram will be two key ingredients allowing to derive theoretical characterization of the resulting tests. The sparsity of the target signals under \mathcal{H}_1 in the Fourier space also allows to adapt testing procedures designed for sparse signals, but not necessarily for periodicity detection.

Some of the fundamental questions at the origin of this work and solved in the next chapters, are: If such reliable simulations of the colored noise under \mathcal{H}_0 are available, how can they be incorporated in the detection process to make it robust ? How reliable can be the significance level associated to an extrasolar planet detection claim when using such simulations ? Which statistical tests are the most appropriate for the detection of the RV Keplerian signatures ? Can we extend detection methods to unevenly sampled time series ?

In Chap. 3, we will propose and analyze different detection methods for the problem of detecting sinusoids with unknown parameters in unknown colored noise. We will consider first the case of regularly sampled observations. The case of an irregular time sampling, encountered in practice in RV, will be treated in Chap. 4.

Part II

Proposed detection method

Chapter 3

Statistical tests exploiting simulations of the colored noise: theoretical analysis and numerical study

Contents

3.1	Statement, assumptions and objectives of the detection problem . . .	82
3.1.1	Model under both hypotheses	82
3.1.2	Assumptions and objectives	83
3.2	Statistical distribution of the proposed standardized periodogram .	84
3.2.1	Standardized periodogram $\tilde{\mathbf{P}} \overline{\mathbf{P}}_L$	84
3.2.2	Statistics of $\tilde{\mathbf{P}} \overline{\mathbf{P}}_L$ under \mathcal{H}_0	84
3.2.3	Statistics of $\tilde{\mathbf{P}} \overline{\mathbf{P}}_L$ under \mathcal{H}_1	85
3.2.4	Mean and variance of $\tilde{\mathbf{P}} \overline{\mathbf{P}}_L$ under \mathcal{H}_0	86
3.3	Statistical tests applied to the standardized periodogram	86
3.3.1	Test of the maximum periodogram value (T_M)	87
3.3.2	Fisher's test and its variations ($T_F, T_{Ch}, T_{F,rob}$)	92
3.3.3	Testing the N_C^{th} largest periodogram value (T_C)	93
3.3.4	Adaptive tests (HC, BJ)	95
3.4	Standardization using parametric estimates of S_E	95
3.4.1	AR parameters estimation	95
3.4.2	Distribution of $\tilde{\mathbf{P}} \hat{\mathbf{S}}_{E,AR}$	97
3.5	Numerical studies on synthetic noise processes	98
3.5.1	Classical tests based on $\tilde{\mathbf{P}} \overline{\mathbf{P}}_L$	99
3.5.2	Adaptive tests based on $\tilde{\mathbf{P}} \overline{\mathbf{P}}_L$	104
3.5.3	Effect of ignoring noise correlations	109
3.5.4	Comparison of the proposed standardization with a parametric approach	111
3.5.5	Application to the design of observational strategies	114
3.6	Summary and conclusions	115

This chapter focuses on the design and statistical characterization of detection tests when the PSD of the colored noise is partially unknown and the signal parameters are unknown.

Following the review of the literature presented in Chap. 2, we propose a standardized periodogram of the form of (2.75). The standardization is obtained by means of an averaged periodogram computed from, say, $L \ll N$ training time series. This training data set is composed of L independent realizations of the noise process alone. One focus of the study is on the gain that can be expected by using such training signals in a detection framework.

We, of course, make the important assumption that such a training data set is available. Beyond the case of exoplanet detection considered here, one may imagine various situations where training signals of the noise can be obtained. In astronomical instruments for instances, secondary optical paths are often devoted to monitor “empty” regions of the sky or calibration stars, in order to provide samples of the noise sources alone (a typical case is in the detection of stellar oscillations [Gupta et al., 2001]). Note that training noise vectors are routinely used for detection in radar systems, with however, an important difference: adaptive test statistics in radar typically use estimates of the covariance matrix of the training vectors and therefore require $L > N$ for this matrix to be nonsingular [Van Trees, 2002]. This is a very different regime from that considered here, where $L \ll N$.

Thus, the proposed statistical approach and the corresponding performance analysis is general and can be useful for a large variety of detection problems, as soon as it is possible to acquire time series of the colored noise.

The results presented in this chapter have been published in [Sulis et al., 2015, Sulis et al., 2016b, Sulis et al., 2017b].

3.1 Statement, assumptions and objectives of the detection problem

3.1.1 Model under both hypotheses

According to the analyzes made in Chap. 2 (cf. Sec. 2.7), we consider the general hypothesis testing model defined as:

$$\begin{cases} \mathcal{H}_0 : X(t) = E(t) & (3.1.a) \\ \mathcal{H}_1 : X(t) = \sum_{q=1}^{N_S} \alpha_q \sin(2\pi f_q t + \varphi_q) + E(t) & (3.1.b) \end{cases} \quad (3.1)$$

where $X(t)$ is an evenly sampled data time series of length N .

Under \mathcal{H}_0 , the noise $E(t)$ is a zero-mean second-order stationary correlated Gaussian noise with PSD such as $\inf(S_E(\nu)) > 0$ and autocorrelation function r_E such as $\sum_k |r_E(k)| < \infty$. This essentially means the noise process has no reason to be band-limited nor coherent in the long term. In our application to exoplanet detection, this noise corresponds to the stochastic process generated by stellar convection and instrumental noise.

Under \mathcal{H}_1 , the target signal is modelled by an unknown number $N_S \geq 1$ of harmonics functions of unknown amplitudes ($\alpha_q \in \mathbb{R}^{*+}$), frequencies ($f_q \in \mathbb{R}^{*+}$) and phases ($\varphi_q \in [0, 2\pi[$). Strictly speaking, as any natural frequency present in the target signal has zero probability to coincide with a Fourier frequency, the corresponding number of nonzero Fourier coefficients (i.e. the number of deviations in the Fourier spectrum under the alternative) always equals N (cf. Fig. 2.9). However, most of the energy of (quasi-) periodic signals is captured by a small fraction of Fourier coefficients, so that model (3.1) would often be accurate for any (quasi) periodic

signals. In the particular case of RV planetary signatures, the study of the influence of the Keplerian parameters, in Sec. 2.3, showed that this is indeed the case. We have seen that, in all cases except perhaps very rare and exotic systems, the RV spectral signatures (1.3) exhibit only a small fraction of significant harmonics because the planets tend to have low eccentricities and are in small number ($N_p \ll N$).

In short, the spectrum is sparse (though not strictly sparse) and RV signals can be modelled by a sum of a small number of pure sinusoids. We call this number N_S and we say that $N_S \ll N$. When there is one planet with frequency close to the Fourier grid, N_S will be essentially 1. In the simulations presented in Sec. 3.4, we will see that multi-planetary systems with 5 eccentric planets behave like model (3.1) with N_S not exceeding, say, 20 at most.

3.1.2 Assumptions and objectives

The working hypotheses made in this chapter are:

1. **Unbiased noise simulations are available.** In the present study, we assume, and this is the originality of the present work, that independent realizations of the colored noise can be generated (e.g. through stellar HD simulations) and consequently that the noise power spectral density, S_E , is partially known through these simulations. This training data set is composed with L time series $\{X_\ell\}_{\ell=1,\dots,L}$ sampled on the same grid as the observations. This set, denoted by \mathcal{T} , is unbiased in the sense that an averaged periodogram of the form of (2.48) obtained from an infinitely large batch would converge uniformly to the true noise PSD. In practice, finite (possibly small) batch sizes can be encountered. For this reason, we say above that the noise is partially unknown, we consider $L \ll N$ and we address the impact of small values of L on the detection performance. In our application example, the simulations of the stellar convection used in [Bigot et al., 2011] are computationally demanding. Obtaining a simulation of 100 days, for a Sun-like star, takes about 3 months of computing time on 120 cores on modern clusters. Consequently, realistic values of L are typically of a few tens at most. The realism of this hypothesis will be addressed in the perspective part through the comparison of MHD simulations with real solar observations. The effect of other noises is overlooked. The effects of coherent stellar oscillations and of spots/plages at the stellar surface are ignored here. Both would lead to quasi-coherent periodic perturbations, which would then be captured by the model under \mathcal{H}_1 . In addition, while the former effect can be accurately reproduced by HD simulations, the latter can not (cf. Sec. 1.4.) Consequently, the results of these thesis attempt to show what we can do, at best, if they are not such kinds of noise sources (spots and pulsations).
2. **The time series $X(t)$ is evenly sampled.** Because one important objective of this study is to obtain analytical characterization of the test performance, we consider first a regular sampling. The case of an irregular sampling, which is important in practice for astronomical applications, will be treated in Chap. 4 (through different techniques as the direct statistical analysis of periodograms is not achievable in this case).
3. **The number of observations, N , is large.** The analytical results obtained in this study are asymptotic in the number of samples, which is characteristic of time series analysis. We will however pay attention to whether asymptotic theory accurately describes reality for finite sample sizes through numerical simulations.

3.2 Statistical distribution of the proposed standardized periodogram

3.2.1 Standardized periodogram $\tilde{P}|\bar{P}_L$

Regarding the problem of controlling the distribution of tests based on the periodogram ordinates in case of unknown noise statistics (cf. Sec. 2.6), we propose a standardization of the periodogram related to (2.75). As seen in Sec. 2.2.8, a straightforward, consistent and asymptotically unbiased estimate of the noise PSD can be obtained by the averaged periodogram obtained from L independent realizations under \mathcal{H}_0 . Replacing \hat{S}_E in (2.75) by \bar{P}_L in (2.48), this standardized periodogram writes:

$$\tilde{P}(\nu_k) := \frac{P(\nu_k)}{\bar{P}_L(\nu_k)}. \quad (3.2)$$

We will sometimes write the standardized periodogram under the form of (3.2) as $\tilde{P}|\bar{P}_L$ when the specification is necessary, i.e. when comparing with a different standardization (e.g. $\tilde{P}|\hat{S}_E$ for a standardization by a specific estimate \hat{S}_E .)

One advantage of using such a periodogram is that the ratio ordinates at different frequencies are asymptotically independent, leading to tractable analytical derivations.

The term ‘‘standardized periodogram’’ has been already used in the past but for different concerns (e.g. to compare the ratio of periodogram ordinates for reflection symmetry [Lu and Zimmerman, 2005] or to calibrate P by some specific estimators of the noise spectrum [Li, 2014]). Such versions of the standardized periodogram differ essentially by the fact that their ‘‘calibrator’’ (the denominator in (3.2)) is based on the observed data while \bar{P}_L is based on independent noise realizations.

3.2.2 Statistics of $\tilde{P}|\bar{P}_L$ under \mathcal{H}_0

Using (2.35), the asymptotic distribution of the averaged periodogram can be derived as [Bartlett, 1950]:

$$\bar{P}_L(\nu|\mathcal{H}_0) \sim \begin{cases} S_E(\nu_k) \frac{\chi_{2L}^2}{2L}, & \forall k \in \Omega \\ S_E(\nu_k) \frac{\chi_L^2}{L}, & \text{for } k = 0, \frac{N}{2}. \end{cases} \quad (3.3)$$

As the numerator and denominator of (3.2) are asymptotically independent variables with known asymptotic distributions, the distribution of their ratio can be deduced. Using (2.35) and (3.3), the asymptotic distribution of \tilde{P} under \mathcal{H}_0 is a scaled ratio of two independent χ^2 distributed r.v., which is a F distribution (cf. Definition. 2.1.2, Sec. 2.1.1):

$$\tilde{P}(\nu_k|\mathcal{H}_0) \sim \begin{cases} \frac{S_E(\nu_k) \chi_2^2/2}{S_E(\nu_k) \chi_{2L}^2/2L} \sim \frac{\chi_2^2/2}{\chi_{2L}^2/2L} \sim F(2, 2L), & \forall k \in \Omega \\ \frac{S_E(\nu_k) \chi_1^2}{S_E(\nu_k) \chi_L^2/L} \sim \frac{\chi_1^2}{\chi_L^2/L} \sim F(1, L), & \text{for } k = 0, \frac{N}{2}. \end{cases} \quad (3.4)$$

This distribution is independent of the noise PSD under \mathcal{H}_0 because the noise influence in the data (numerator) is cancelled by the noise training data set (denominator). This key property will make the detection robust, as constant false alarm rates can be guaranteed whatever S_E (recall that S_E is not exactly known).

One can note that a F distribution was also derived and exploited in [Lu and Zimmerman, 2005] for symmetry testing purposes (ratios of the form $P(\nu_k)/P(\nu_l)$, $k \neq l$) assuming the periodogram components are asymptotically independent.

3.2.3 Statistics of $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$ under \mathcal{H}_1

The distribution of the periodogram P is known when cisoids are present under the alternative of model (3.1) (cf. (2.39)). The real case of model (3.1) is slightly more tedious but can be treated similarly and gives a non central F distribution (cf. Appendix. B)

$$\tilde{P}(\nu_k|\mathcal{H}_1) \sim \begin{cases} \frac{S_E(\nu_k)\chi_{2,\lambda_k}^2}{S_E(\nu_k)\chi_{2L}^2} \sim \frac{\chi_{2,\lambda_k}^2/2}{\chi_{2L}^2/2L} \sim F_{\lambda_k}(2, 2L), & \forall k \in \Omega, \\ \frac{S_E(\nu_k)\chi_{1,\lambda_k}^2}{S_E(\nu_k)\chi_L^2/L} \sim \frac{\chi_{1,\lambda_k}^2}{\chi_L^2/L} \sim F_{\lambda_k}(1, L), & \text{for } k = 0 \text{ and } k = \frac{N}{2}. \end{cases} \quad (3.5)$$

where the non centrality parameter λ_k depends on the signal signatures. Adapting the proof of Theorem 6.2 of [Li, 2014], one can show that, $\forall k \in \Omega$:

$$\lambda_k = \rho_k(r_E) \frac{N}{2S_E(\nu_k)} \sum_{q=1}^{N_S} \left[\alpha_q^2 \kappa_q^2 + 2\alpha_q \kappa_q \sum_{\ell=q+1}^{N_S} \alpha_\ell \kappa_\ell \cos(\theta_q - \theta_\ell) \right], \quad (3.6)$$

and for $k = 0, \frac{N}{2}$ this expression is halved. The proof of (3.6) is reported in Appendix. B.

In general, the term $\rho_k(r_E)$ in (3.6), given in (2.36), can be approximated to 1 as the noise autocorrelation function r_E vanishes for large N (cf. Sec. 2.2.6). This approximation will be studied numerically in Sec. 3.5.1 for different correlated noises.

The terms κ_q and θ_q in (3.6) arise from the spectral leakage of sinusoids with off-grid frequencies (through the spectral window K_N defined in (2.27)). These values are given by the modulus and argument of the complex variable z_q :

$$z_q(\nu_k) := D_N(f_q - \nu_k) e^{i(\varphi_q - \frac{\pi}{2})} - D_N(f_q + \nu_k) e^{-i(\varphi_q + \frac{\pi}{2})}, \quad (3.7)$$

$$\begin{aligned} \kappa_q &:= |z_q(\nu_k)| \\ &= (K_N(f_q - \nu_k)^2 + K_N(f_q + \nu_k)^2 - 2K_N(f_q + \nu_k)K_N(f_q - \nu_k)\cos(2\pi(N+1)f_q + 2\varphi_q))^{1/2} \end{aligned} \quad (3.8)$$

$$\theta_q := \angle z_q, \quad \theta_q \in] -\pi, \pi]. \quad (3.9)$$

The non-centrality parameter λ_k in (3.6) is evaluated at the Fourier frequencies ν_k and corresponds to the signal power (plus an interference term) normalized by the noise power in the k -th frequency bin. Expression (3.6) is composed of a sum involving the signal parameters $\boldsymbol{\theta}_q = [\alpha_q, f_q, \varphi_q]^\top$, for $q = 1, \dots, N_S$, and the colored noise PSD S_E . The finite duration of the time series and the presence of off-grid frequency signals leads to spectral leakage, characterised by the Fejér kernel K_N (or spectral window, cf. Sec. 2.2) in the κ_q terms. This kernel is composed of a main central lobe and sidelobes, whose widths depend on the spectral resolution ($\frac{1}{N\Delta t}$) and on the signal frequencies position on the Fourier grid. The magnitude of these sidelobes at neighbouring frequencies can be significant when the signal is off-grid and the superposition of all the ‘‘signals kernel sidelobes’’ are evaluated by the second sum in (3.6). In the case where the signal frequencies are on the Fourier grid (i.e. model (3.1).b is exact), the second sum tends to zero as $K_N(0) = 1$ (leading to $\kappa_q(f_q) \rightarrow 1$ and $\kappa_q(\nu_k \neq f_q) \rightarrow 0, \forall \nu_k$, and so a vanishing product $\kappa_q \kappa_\ell$). In this case, equation (3.6) is equivalent to the expression given in (2.40) with $\kappa_q = |D_N(\nu_k - f_j)|^2$ (cf. remark 6.6 of [Li, 2014]).

As already discussed in Sec. 3.1.1, for general (possibly quasi-) periodic target signatures, the decomposition into N_S harmonic frequencies $\{f_q \in \mathbb{R}^{*+}\}$ in model (3.1) is inaccurate. So the number of non zero parameters λ_k is N . This is also the case when model (3.1).b is exact with at least one frequency f_s not on the Fourier grid, even if $N_S \ll N$. However, owing to the efficiency of Fourier representation for compacting energy and to the fast decay of $K_N(\nu)$ as a function of ν , the proportion of parameters λ_k that significantly differ from 0 is always small w.r.t. N .

3.2.4 Mean and variance of $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$ under \mathcal{H}_0

Using (3.5), it is interesting to evaluate the mean and variance of the r.v. $\tilde{P}|\bar{P}_L$. Using the theoretical expressions for a F distributed r.v. given in Table. 2.2, we obtain

$$\begin{cases} \mathbb{E}[\tilde{P}(\nu)] = \frac{2L}{2L-2} = \frac{L}{L-1}, & \forall L > 1 \\ \text{Var}[\tilde{P}(\nu)] = \frac{2(2L)^2(2+2L-2)}{2(2L-2)^2(2L-4)} = \frac{L^3}{(L-1)^2(L-2)}, & \forall L > 2. \end{cases} \quad (3.10)$$

These quantities are only dependent on the number of available training data set (L). The mean is only defined for $L > 1$ and the variance for $L > 2$. Consequently, one can expect unpredictable deviations when only one or two training time series are available to standardize the periodogram in (3.2). The reason is that the noise PSD is not well constrained due to the high variance of \bar{P}_L making the standardization unstable. In the case of large L ($L \rightarrow \infty$), the mean tends to 1 because \bar{P}_L is an unbiased estimate of P . The variance also tends to 1 as $\frac{\text{Var}[P]}{S_E} = \frac{S_E}{S_E} = 1$ when $L \rightarrow \infty$.

Example 12. Numerical illustration of the mean and variance of $\tilde{P}|\bar{P}_L$

Panel a) of Fig. 3.1 illustrates the theoretical mean (red) and variance (blue) given in (3.10) as a function of L . The horizontal black dashed line represents the expected asymptotic value for these quantities (i.e. 1). One can observe that this asymptotic value is achieved very fast (when L reaches a few tens).

The left panels compare these quantities with empirical results obtained by generating 10^4 standardized periodograms by MC simulations with b) $L = 1$, and c) $L = 100$ noise training time series. For computing $\mathbb{E}[\tilde{P}(\nu)]$ and $\text{Var}[\tilde{P}(\nu)]$, we use the time series of the AR(2) process introduced in Example. 8 with $N = 1000$. As expected, one observes occurrences of large deviations to the asymptotic value in the case of low L , while this does not happen for larger L values.

3.3 Statistical tests applied to the standardized periodogram

To simplify the presentation of the results, one will restrict in the following to the frequency set Ω , i.e., to standardized vectors

$$\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L := \left[\frac{P(\nu_1)}{\bar{P}_L(\nu_1)}, \dots, \frac{P(\nu_{\frac{N}{2}-1})}{\bar{P}_L(\nu_{\frac{N}{2}-1})} \right]^\top,$$

with $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$ the standardized periodogram obtained by means of an averaged periodogram of L training data set of the colored noise. Extension of the results to all frequencies is straightforward using the distributions given in (3.4) and (3.5).

Several statistical tests can be applied to the ordinates of this standardized periodogram. As those ordinates are asymptotically mutually independent and further independent on the noise PSD, with known distribution, the false alarm and detection probabilities can be derived for some of these tests and some of them turn out to be CFAR. The following subsections derive the statistical properties of the main tests studied in Sec. 2.4 when applied to $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$.

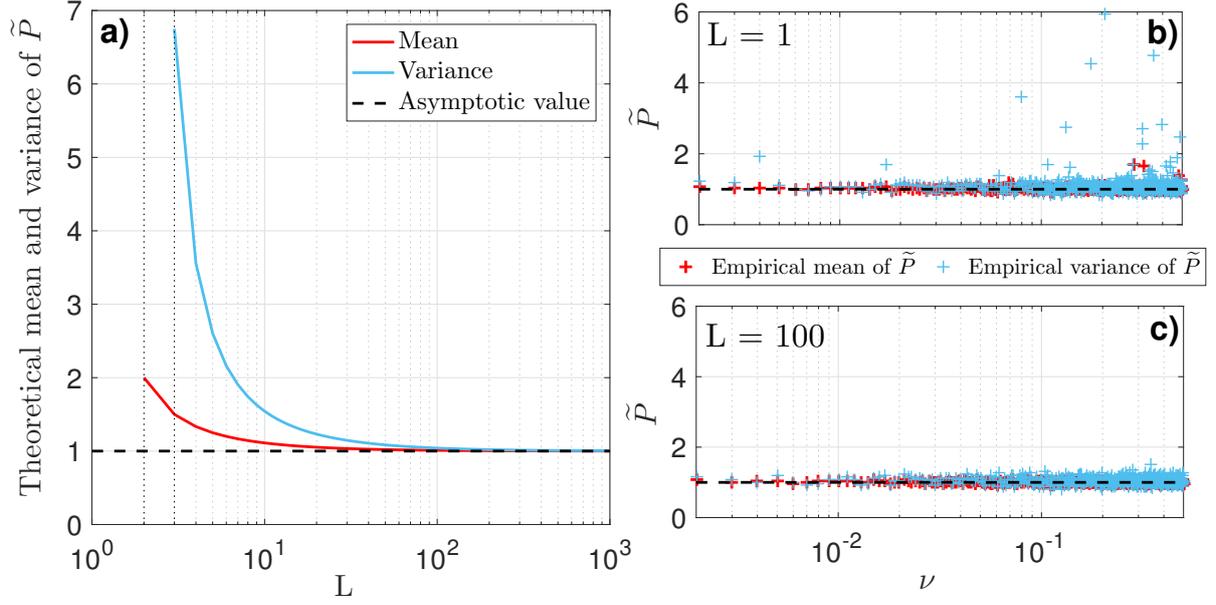


Figure 3.1 – Mean (red) and variance (blue) of the standardized periodogram. a) Theoretical behavior as a function of L according to (3.10). Numerical evaluation of (3.10) using 10^4 MC simulations of an AR(2) process for b) $L = 1$, and c) $L = 100$ training data sets. In all panels, the black horizontal dashed line represents the asymptotic value (1).

3.3.1 Test of the maximum periodogram value (T_M)

First, consider the most famous test (almost always used when looking for extrasolar planet signatures in RV data, e.g. [Scargle, 1982, Cumming, 2004, Zechmeister and Kürster, 2009]), the test of the maximum periodogram value, given in (2.51). When applied to the standardized periodogram, this test writes:

$$T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L) := \max_{\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\geq}} \gamma. \quad (3.11)$$

with $\gamma > 0$.

Under \mathcal{H}_0 of model (3.1) and in the asymptotic regime, this test statistics is the maximum of independent and identically F distributed r.v., cf. (3.4). Using the probability density function φ_F of a F r.v. given in (2.6), the density of a $F(2, 2L)$ r.v. writes

$$\begin{aligned} \varphi_F(\gamma, 2, 2L) &= \frac{\sqrt{\frac{(2\gamma)^2(2L)^{2L}}{(2\gamma+2L)^{2+2L}}}}{\gamma B\left(\frac{2}{2}, \frac{2L}{2}\right)} \\ &= \frac{1}{\gamma B(1, L)} \frac{\gamma 2^{L+1} L^L}{2^{L+1}(\gamma + L)^{L+1}}, \\ &= \frac{1}{B(1, L)} \frac{1}{L} \left(\frac{L}{\gamma + L}\right)^{L+1}. \end{aligned} \quad (3.12)$$

Using the Beta function definition in (2.7), the term $B(1, L)$ writes

$$B(1, L) := \int_0^1 (1-t)^{L-1} dt = \frac{1}{L}. \quad (3.13)$$

By replacing $B(1, L)$ in (3.12), the density $\varphi_F(\gamma, 2, 2L)$ becomes

$$\varphi_F(\gamma, 2, 2L) = \frac{1}{B(1, L)} \cdot \frac{1}{L} \cdot \left(1 + \frac{\gamma}{L}\right)^{-L-1} = \left(1 + \frac{\gamma}{L}\right)^{-L-1}.$$

It can be checked that the pdf is indeed normalized:

$$\int_0^\infty \varphi_F(\gamma, 2, 2L) d\gamma = 1.$$

Using now the cdf of a $F(2, 2L)$ r.v. given in (2.8), the corresponding CDF $\Phi_F(\gamma, 2, 2L)$ writes

$$\Phi_F(\gamma, 2, 2L) = I_{\frac{2\gamma}{2\gamma+2L}}\left(\frac{2}{2}, \frac{2L}{2}\right), \quad (3.14)$$

with $I_x(a, b)$ the regularized incomplete beta function ($x, a, b \in \mathbb{R}^+$), defined in (2.9). As parameters $a, b \in \mathbb{N}$, one can use integration by parts to compute I . By further developing the integral in (2.9), function $I_x(a, b)$ can be rewritten in terms of binomial expansion as (cf. eq. 6.6.4, p.263 of [Abramowitz et al., 1972]):

$$I_x(a, b) = \sum_{j=a}^{a+b-1} \frac{(a+b-1)!}{j!(a+b-1-j)!} x^j (1-x)^{a+b-1-j}. \quad (3.15)$$

Finally, using (3.15) and the Newton's binomial formula:

$$(a+b)^n = \sum_{i=0}^n C_n^i a^i b^{n-i}, \quad (3.16)$$

the cdf of $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ in (3.14) can be developed in a remarkably simple analytical expression:

$$\begin{aligned} \Phi_F(\gamma, 2, 2L) &:= I_{\frac{\gamma}{L+\gamma}}(1, L) \\ &= \sum_{j=1}^L \binom{L}{j} \left(\frac{\gamma}{L+\gamma}\right)^j \left(1 - \frac{\gamma}{L+\gamma}\right)^{L-j} \\ &= \sum_{j=0}^L \binom{L}{j} \left(\frac{\gamma}{L+\gamma}\right)^j \left(1 - \frac{\gamma}{L+\gamma}\right)^{L-j} - \binom{L}{0} \left(\frac{\gamma}{L+\gamma}\right)^0 \left(1 - \frac{\gamma}{L+\gamma}\right)^{L-0} \\ &= \left(\frac{\gamma}{L+\gamma} + 1 - \frac{\gamma}{L+\gamma}\right)^L - \left(1 - \frac{\gamma}{L+\gamma}\right)^L \\ &= 1 - \left(\frac{L}{L+\gamma}\right)^L. \end{aligned} \quad (3.17)$$

The probability of false alarm, defined in (2.4), of test $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ in (2.51) can be computed thanks to the asymptotic independence of the $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$ ordinates :

$$\begin{aligned} P_{FA}(T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L), \gamma) &:= \Pr(T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L) > \gamma | \mathcal{H}_0) \\ &= 1 - \prod_{k \in \Omega} \Pr(\tilde{P}(\nu_k) \leq \gamma | \mathcal{H}_0) \\ &= 1 - \left(\Phi_F(\gamma, 2, 2L)\right)^{\frac{N}{2}-1} \\ &= 1 - \left(1 - \left(\frac{L}{\gamma+L}\right)^L\right)^{\frac{N}{2}-1}. \end{aligned} \quad (3.18)$$

The P_{FA} is independent of the noise PSD, which makes $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ a CFAR detector. This property has been developed in the asymptotic regime framework, but as we will show in Sec. 3.5.1, this expression is still reliable for finite and relatively low N values depending on

the correlation timescale ($N = 1024$ in the considered example).

Under \mathcal{H}_1 of model (3.1), using (2.39) and the approximate independence of the periodogram ordinates in (3.2) ([Li, 2014], Theorem 6.5), one can deduce that the r.v. $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ in (3.11) follows a non central F distribution $F_{\lambda_k}(2, 2L)$. Consequently, the probability of detection of $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ can be similarly approximated as:

$$P_{DET}(T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L), \gamma) := \Pr\left(T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L) > \gamma | \mathcal{H}_1\right) \approx 1 - \prod_{k \in \Omega} \Phi_{F_{\lambda_k}}(\gamma, 2, 2L). \quad (3.19)$$

Using (3.18), one can also deduce the threshold expression as a function of the false alarm, i.e. $\gamma(P_{FA})$, for the test statistics T_M :

$$\begin{aligned} 1 - \left(1 - P_{FA}(\gamma)\right)^{\frac{1}{\eta}} &= \left(\frac{L}{\gamma + L}\right)^L \\ \frac{L}{\gamma + L} &= \left(1 - \left(1 - P_{FA}(\gamma)\right)^{\frac{1}{\eta}}\right)^{\frac{1}{L}} \\ \gamma + L &= L \left(1 - \left(1 - P_{FA}(\gamma)\right)^{\frac{1}{\eta}}\right)^{-\frac{1}{L}}, \end{aligned} \quad (3.20)$$

which gives

$$\gamma(T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L), P_{FA}) = L \left[\left(1 - (1 - P_{FA})^{\frac{1}{N/2-1}}\right)^{-\frac{1}{L}} - 1 \right]. \quad (3.21)$$

Finally, using (3.19) and (3.21), one can express the detection probability for a given false alarm, i.e. $P_{DET}(P_{FA})$, which can be useful to compute ROC curves:

$$P_{DET}(T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L), P_{FA}) \approx 1 - \prod_{k \in \Omega} \Phi_{F_{\lambda_k}}\left(L(1 - (1 - P_{FA})^{\frac{1}{\eta}})^{-\frac{1}{L}} - L, 2, 2L\right). \quad (3.22)$$

Illustrations of these results and the reliability of (3.18), (3.19) and (3.22) will be demonstrated by numerical studies in Sec. 3.5.1.

Case of N_S sinus on the grid. If all signal frequencies $\{f_j\}_{j=1, \dots, N_S}$ of model (3.1.b) fall on the Fourier frequency grid, the P_{DET} of $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ given in (3.19) becomes:

$$P_{DET}(T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L), \gamma) = 1 - \left(\Phi_{F(\gamma, 2, 2L)}\right)^{\frac{N}{2}-1-N_S} \prod_{j=1}^{N_S} \Phi_{F_{\lambda_j}(\gamma, 2, 2L)}.$$

with the non centrality parameters given in (3.6) becoming here:

$$\lambda_k = \begin{cases} \rho_j \frac{N}{2} \sum_{j=1}^{N_S} \frac{\alpha_j^2}{S_E(f_j)}, & \text{for } k = j \\ 0, & \text{for } k \neq j \end{cases} \quad (3.23)$$

since for $k = j$, the crossed term in (3.6) vanishes owing to the orthogonality of the Fejér kernels in the κ_j terms (3.2.3). One can note that expression (3.23) reduces to (2.44) in case of WGN, because $S_E(f_j) = \sigma^2$ and $\rho_j = 1, \forall j = 1, \dots, N$ since $r_E(u) = 0, \forall u > 1$.

Analytical characterization of $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ under \mathcal{H}_0 To quantify the behavior of the test statistics $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ as a function of N and L , it is interesting to look at (i) its distribution, (ii) mode, and (iii-iv) first moments under the null.

(i) Density. For large N , the *Extreme value Theorem* predicts that all the maximum of r.v. converges towards an extreme value distribution (cf. [Coles, 2001] and also Sec. 4.3). In the case of $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$, one can develop the density function at finite N using (2.35). Using also the periodogram distribution under the null (2.35), the cdf of $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$, i.e. of the maximum of $\frac{N}{2} - 1$ F independent variables, writes

$$\Phi_{T_M}(\gamma, 2, 2L) := \left[1 - \left(\frac{L}{\gamma + L}\right)^L\right]^{\frac{N}{2}-1} = \left[1 - g(\gamma)^L\right]^{\frac{N}{2}-1} \quad (3.24)$$

with $g(\gamma) := \frac{L}{\gamma + L}$. One can deduce the pdf by integrating (3.24). Using the relation $\frac{df(x)^y}{dx} = y f(x)^{y-1} f'(x)$, this derivative writes

$$\begin{aligned} \varphi_{T_M}(\gamma, N, L) &:= \frac{\partial \Phi_F(\gamma, 2, 2L)}{\partial \gamma} \\ &= \frac{\partial}{\partial \gamma} \left[\left[1 - g(\gamma)^L\right]^{\frac{N}{2}-1} \right] \\ &= \left(\frac{N}{2} - 1\right) \left[1 - g(\gamma)^L\right]^{\frac{N}{2}-2} \frac{\partial}{\partial \gamma} \left[1 - g(\gamma)^L\right] \\ &= \left(\frac{N}{2} - 1\right) \left[1 - g(\gamma)^L\right]^{\frac{N}{2}-2} (-L) g(\gamma)^{L-1} \left((-1)g(\gamma)^2 \frac{1}{L}\right) \\ &= \left(\frac{N}{2} - 1\right) \left[1 - g(\gamma)^L\right]^{\frac{N}{2}-2} g(\gamma)^{L-1} g(\gamma)^2 \\ &= \left(\frac{N}{2} - 1\right) \left[1 - g(\gamma)^L\right]^{\frac{N}{2}-2} g(\gamma)^{L+1} \\ &= \left(\frac{N}{2} - 1\right) \left[1 - \left(\frac{L}{\gamma + L}\right)^L\right]^{\frac{N}{2}-2} \left(\frac{L}{\gamma + L}\right)^{L+1}. \end{aligned} \quad (3.25)$$

(ii) Mode. The mode is derived at the point cancelling the pdf derivative, say $\gamma = \gamma^* > 0$:

$$\begin{aligned} \left. \frac{d\varphi_{T_M}(\gamma; N, L)}{d\gamma} \right|_{\gamma=\gamma^*} &= \frac{d}{d\gamma} \left[\left(\frac{N}{2} - 1\right) \left[1 - g(\gamma)^L\right]^{\frac{N}{2}-2} g(\gamma)^{L+1} \right]_{\gamma=\gamma^*} \\ &= \left(\frac{N}{2} - 1\right) \frac{d}{d\gamma} \left[\left[1 - g(\gamma)^L\right]^{\frac{N}{2}-2} g(\gamma)^{L+1} \right]_{\gamma=\gamma^*} \\ &= \left(\frac{N}{2} - 1\right) \left[\left(\frac{N}{2} - 2\right) \left[1 - g(\gamma)^L\right]^{\frac{N}{2}-3} g(\gamma)^{2(L+1)} - \frac{L+1}{L} g(\gamma)^{L+2} \left[1 - g(\gamma)^L\right]^{\frac{N}{2}-2} \right]_{\gamma=\gamma^*} \\ &= 0, \end{aligned} \quad (3.26)$$

(iii) First moment. The esperance of $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ is defined for $N, L > 4$:

$$\mathbb{E} \left[\max_k \tilde{\mathbf{P}} \right] = \int_0^\infty \gamma \varphi(\gamma, N, L) d\gamma = \left(\frac{N}{2} - 1\right) \int_0^\infty \gamma \left[1 - g(\gamma)^L\right]^{\frac{N}{2}-2} g(\gamma)^{L+1} d\gamma \quad (3.27)$$

(iv) The second moment.

$$\text{Var} \left[\max_k \tilde{\mathbf{P}} \right] = \int_0^\infty \gamma^2 \varphi(\gamma, N, L) d\gamma = \left(\frac{N}{2} - 1\right) \int_0^\infty \gamma^2 \left[1 - g(\gamma)^L\right]^{\frac{N}{2}-2} g(\gamma)^{L+1} d\gamma \quad (3.28)$$

As (3.26), (3.27) and (3.28) are not derivable analytically, these equations have to be resolved numerically (e.g. using the global adaptive quadrature integrator of MATLAB, cf. [Shampine, 2008]).

Fig. 3.2 illustrates the influence of the number of observations N (left, with L fixed to 20) and of the number of noise training data set L (right, with N fixed to 1000) on the distribution of $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ under \mathcal{H}_0 given in (3.25). One can observe the increase of the pdf mode with N but this effect will be more significant under the alternative (not shown, cf. Sec. 2.2.8). One also sees that the pdf becomes more narrow for larger L . When $L \rightarrow \infty$, i.e. when the noise PSD is perfectly estimated by the averaged periodogram, the standardized periodogram in (3.2) is χ_2^2 distributed.

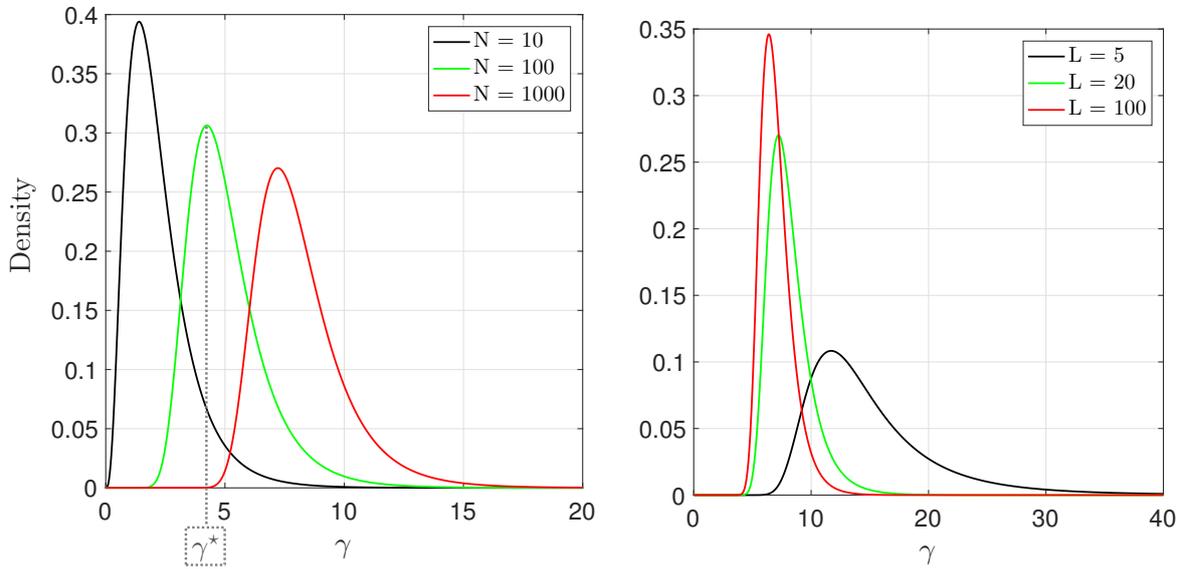


Figure 3.2 – Density of $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ given in (3.25) as a function of the detection threshold for various values of N (left, for fixed $L = 20$) and L (right, for fixed $N = 1000$). The value $\gamma = \gamma^* > 0$ indicates the mode of the distribution for $L = 20$ and $N = 100$.

Fig. 3.3 represents (resp. from left to right) the mode and the two first moments of $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ given in (3.26), (3.27) and (3.28) as a function of N (first row, fixed $L = 20$) and L (second row, fixed $N = 1000$). One value $\gamma = \gamma^*$ localized in Fig. 3.2 is shown just for verification. First, one can observe the increase of all considered quantities with N more rapidly for low L . It illustrates the fact that when more data are available, the test statistics has high value but at the cost of a larger dispersion (especially if L is low, cf. Sec.(3.2.4)). The mode and the two first moments decrease with L (more rapidly for low N). As a consequence, the FA threshold associated to test $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ will decrease for a large number of noise training data set (facilitating then the detection). We see that these decreases are stabilized when L reaches a few tens, meaning that we can expect to reach the asymptotic (in terms of L) FA thresholds using a moderate number of training data set. For large L , one can expect less significant variations of the three quantities with N under the null.

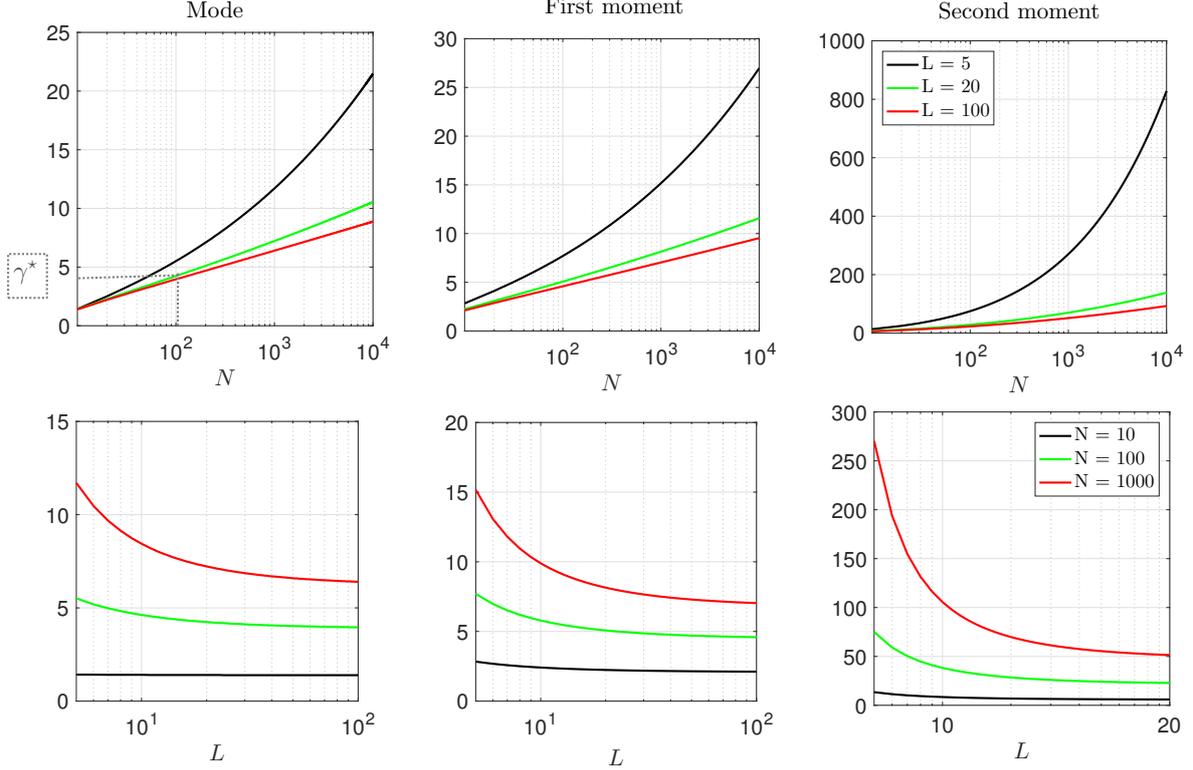


Figure 3.3 – Theoretical mode, esperance and variance of $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ given in (3.26), (3.27) and (3.28) as function of N (first row, fixed $L = 20$) and L (second row, fixed $N = 1000$).

3.3.2 Fisher's test and its variations (T_F , T_{Ch} , $T_{F,rob}$)

Under \mathcal{H}_0 of model (3.1), the Fisher's test (2.59) applied to $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$ becomes:

$$T_F(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L) := \frac{\tilde{P}_{(\frac{N}{2}-1)}}{\sum_{k=1}^{\frac{N}{2}-1} \tilde{P}_{(k)}}. \quad (3.29)$$

This test looks for the maximum of identically distributed variables, which from (3.4) are distributed as:

$$\frac{\frac{P(\nu_i)}{\bar{P}_L(\nu_i)}}{\sum_k \frac{P(\nu_k)}{\bar{P}_L(\nu_k)}} \sim \frac{F(2, 2L)}{\sum_k F(2, 2L)}. \quad (3.30)$$

Each variable is a ratio of a F variable over a sum of F variables.

When $L \rightarrow \infty$, the F r.v. converges in distribution to a χ_2^2 r.v. (i.e. $F(2, 2L) \xrightarrow{d} \chi_2^2/2$) leading to a P_{FA} as in (2.60) (cf. p.324 of [Johnson et al., 1994]).

For finite values of L , no analytic characterization of the resulting distribution of (3.30) is known to our knowledge but, resorting to Monte Carlo simulations to evaluate the function $\gamma \mapsto P_{FA}$ is possible in principle. Resorting to MC simulations, when L is large, consists in generating standardized periodograms using $L \times (\ell + 1)$ simulated times series of the noise under the null (one for the numerator and ℓ for the denominator). Then, the test is applied on the training periodograms, and the P_{FA} is deduced using the empirical distribution function. When L is small however, assessing the distribution of the maximum of these variables by MC

simulations to compute the false alarm rate is not possible.

Similar remarks can be made about standardized versions of T_{Ch} in (2.63) or $T_{F,rob}$ in (2.64), which involve denominators as in (3.30) in their test statistics: when $L \rightarrow \infty$ their distribution tend to a χ_2^2 distribution and one can use MC simulations, but for finite L the false alarm rate cannot be computed.

Under \mathcal{H}_1 of model (3.1), the same arguments show that one cannot control the distributions of the r.v. $T_F(\tilde{P}(\nu_k)|\bar{\mathbf{P}}_L)$, $T_{Ch}(\tilde{P}(\nu_k)|\bar{\mathbf{P}}_L)$ and $T_{F,rob}(\tilde{P}(\nu_k)|\bar{\mathbf{P}}_L)$ in general. Consequently, for the Fisher test and its variations, the associated detection probabilities cannot be evaluated.

Although computing the P_{FA} of these tests is not possible, one can use them for reference to compare their performance w.r.t. to other tests applied to $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$ or \mathbf{P} (cf. Sec. 3.5.1 and Sec. 3.5.3), but keeping in mind that they cannot be implemented in practice.

3.3.3 Testing the N_C^{th} largest periodogram value (T_C)

We turn now to the test statistic T_C given in (2.61). When applied to $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$, it writes:

$$T_C(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L; N_C) := \tilde{\mathbf{P}}_{(N-N_C+1)}, \quad (3.31)$$

where parameter N_C allows to focus on the regions of order statistics where deviations under the alternative are likely to be most significant.

Under \mathcal{H}_0 , owing to (3.4), each ordinate $(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)_i := \frac{\tilde{P}(\nu_i)}{\bar{P}_L(\nu_i)}$ has probability

$$u := 1 - \Phi_F(\gamma, 2, 2L) = \left(\frac{L}{\gamma + L} \right)^L, \quad (3.32)$$

to be larger than γ . Since the $\frac{N}{2} - 1$ ordinates of $\tilde{\mathbf{P}}$ are asymptotically independent, the number K of standardized ordinates larger than γ under \mathcal{H}_0 follows a binomial distribution :

$$K \sim Bin(N/2 - 1, 1 - u). \quad (3.33)$$

The cdf of a binomial r.v., say Φ_{Bin} , is a I function [Abramowitz et al., 1972]:

$$\Phi_{Bin}(x, n, p) := \Pr(Bin(n, p) \leq x) = I_{1-p}(n - x, 1 + x). \quad (3.34)$$

So, replacing parameters (x, n, p) in (3.34) by $(k, \frac{N}{2}, u)$ gives:

$$\Pr(K \leq k) = I_u(N/2 - k, k). \quad (3.35)$$

Using (3.32) and the relations

$$\Pr(Bin(n, p) \leq j) = 1 - I_{1-p}(j + 1, n - j), \quad (3.36)$$

and

$$1 - I_{1-x}(b, a) = I_x(a, b), \quad (3.37)$$

(cf., Prop. 6.6.3 in [Abramowitz et al., 1972]), the P_{FA} of test T_C applied to $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$ can be

expressed as:

$$\begin{aligned}
P_{FA}(T_C(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L, N_C), \gamma) &:= \Pr(T_C(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L, N_C) > \gamma | \mathcal{H}_0) \\
&= 1 - \sum_{k=0}^{N_C-1} \Pr(K = k | \mathcal{H}_0) \\
&= 1 - \Pr(K < N_C), \\
&= 1 - \Phi_{Bin}(u, \frac{N}{2} - N_C, N_C), \quad \text{using (3.33) and (3.35)} \\
&= 1 - I_{1-u}(\frac{N}{2} - N_C, N_C), \quad \text{using (3.36)} \\
&= I_u(N_C, \frac{N}{2} - N_C), \quad \text{using (3.37)}.
\end{aligned} \tag{3.38}$$

As a side remark, note that the false alarm rate of test T_C applied to the periodogram of a white noise of known variance was obtained as above in (2.62) (cf. Sec. 2.4.3) by replacing Φ_F with the CDF of χ_2^2 r.v. according to (2.35).

As for $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$, Eq.(3.38) shows that test $T_C(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ is CFAR.

Under \mathcal{H}_1 , the function $\gamma \mapsto P_{DET}(\gamma)$ of $T_C(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L, N_C)$ can be deduced similarly to (3.38) with, as we shall see, the difference that K is no longer binomial owing to the non centrality parameter $\{\lambda_k\}$.

Denoting by K the number of ordinates of $\tilde{\mathbf{P}} | \bar{\mathbf{P}}_L$ larger than γ under \mathcal{H}_1 and by $p_i := \Pr(K = i | \mathcal{H}_1)$, the detection probability writes:

$$\begin{aligned}
P_{DET}(T_C(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L, N_C), \gamma) &:= \Pr(T_C(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L, N_C) > \gamma | \mathcal{H}_1) \\
&= \Pr(K \geq N_C | \mathcal{H}_1) \\
&= 1 - \sum_{i=0}^{N_C-1} \Pr(K = i | \mathcal{H}_1) \\
&= 1 - \sum_{i=0}^{N_C-1} p_i.
\end{aligned} \tag{3.39}$$

Owing to (3.5) each ordinate $(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)_i := \frac{\tilde{P}(\nu_i)}{\bar{P}_L(\nu_i)}$ has probability $1 - \Phi_{F_{\lambda_i}}(\gamma)$ with λ_i given in (3.6) to be larger than γ . These variates can be considered approximately independent (asymptotically) but obviously not identically distributed. This is why the variable K is not binomially distributed (as it is under \mathcal{H}_0) and the probabilities $\{p_i\}$ require further investigation. We proceed by induction. In the following, all probabilities are under \mathcal{H}_1 . The first probability p_0 can simply be approximated as

$$p_0 = \Pr \left\{ \bigcap_{k=1}^{\frac{N}{2}-1} (\tilde{\mathbf{P}} | \bar{\mathbf{P}}_L)_k \leq \gamma \right\} \approx \prod_{k=1}^{\frac{N}{2}-1} \Phi_{F_{\lambda_k}}.$$

The probability $p_1 = \Pr(K = 1)$ is similarly

$$p_1 = \Pr \bigcup_{k=1}^{\frac{N}{2}-1} \left\{ (\tilde{\mathbf{P}} | \bar{\mathbf{P}}_L)_k > \gamma \bigcap_{j \neq k} (\tilde{\mathbf{P}} | \bar{\mathbf{P}}_L)_j \leq \gamma \right\} \approx \sum_{k=1}^{\frac{N}{2}-1} \left[(1 - \Phi_{F_{\lambda_k}}) \prod_{\substack{j=1, \\ j \neq k}}^{\frac{N}{2}-1} \Phi_{F_{\lambda_j}} \right].$$

To generalize further, denote by $\Omega^{(i)}$ one particular combination of i indices taken in Ω and $\bar{\Omega}^{(i)} := \Omega \setminus \Omega^{(i)}$ the set of remaining indices. Let $\{\Omega_1^{(i)}, \dots, \Omega_i^{(i)}\}$ (resp. $\{\bar{\Omega}_1^{(i)}, \dots, \bar{\Omega}_{\frac{N}{2}-1-i}^{(i)}\}$) denote the indices in two such combinations, and let Ω^i be the set of all the $\{\Omega^{(i)}\}$. With these notations, one obtains for $i > 1$:

$$p_i = \Pr \bigcup_{\Omega^{(i)} \in \Omega^i} \left\{ \bigcap_{k=1}^i (\tilde{\mathbf{P}} | \bar{\mathbf{P}}_L)_{\Omega_k^{(i)}} > \gamma \bigcap_{k'=1}^{\frac{N}{2}-1-i} (\tilde{\mathbf{P}} | \bar{\mathbf{P}}_L)_{\bar{\Omega}_{k'}^{(i)}} \leq \gamma \right\} \quad (3.40)$$

$$\approx \sum_{\Omega^{(i)}} \prod_{k=1}^i \left(1 - \Phi_{F_{\lambda_{\Omega_k^{(i)}}}(\gamma, 2, 2L)} \right) \prod_{k'=1}^{\frac{N}{2}-1-i} \Phi_{F_{\lambda_{\bar{\Omega}_{k'}^{(i)}}}(\gamma, 2, 2L)}.$$

The final $P_{DET}(T_C(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L, N_C), \gamma)$ expression follows by combining (3.39) and (3.40) and gives:

$$P_{DET}(T_C(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L, N_C), \gamma) := \Pr(T_C(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L, N_C) > \gamma | \mathcal{H}_1)$$

$$\approx 1 - \sum_{i=0}^{N_C-1} \sum_{\Omega^{(i)} \in \Omega^i} \prod_{k=1}^i \left(1 - \Phi_{F_{\lambda_{\Omega_k^{(i)}}}(\gamma, 2, 2L)} \right) \prod_{k'=1}^{\frac{N}{2}-1-i} \Phi_{F_{\lambda_{\bar{\Omega}_{k'}^{(i)}}}(\gamma, 2, 2L)}, \quad (3.41)$$

which is typographically heavy but can be used to compute ROC curves with (3.38) and (3.41). The non centrality parameters $\{\lambda_{\Omega_k^{(i)}} := \lambda(\nu_{\Omega_k^{(i)}})\}$ and $\{\lambda_{\bar{\Omega}_{k'}^{(i)}} := \lambda(\nu_{\bar{\Omega}_{k'}^{(i)}})\}$ are given by (3.6).

3.3.4 Adaptive tests (HC, BJ)

When applied to $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$, the P -values involved in the Higher Criticism (2.70) and the Berk Jones tests (2.74) should be computed according to the distribution under the null given by (3.4). Hence, similarly to (2.73), the P -values of $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$ at each frequency is:

$$V_{\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L}(k) := \begin{cases} 1 - \Phi_F \left(\frac{P(\nu_k)}{\bar{P}_L(\nu_k)}, 2, 2L \right), & \forall k \in \Omega \\ 1 - \Phi_F \left(\frac{P(\nu_k)}{\bar{P}_L(\nu_k)}, 1, L \right), & \text{for } k = 0, \frac{N}{2}. \end{cases} \quad (3.42)$$

The properties of the adaptive tests HC and BJ are otherwise left unchanged, with the CFAR property added: thanks to the standardization of \mathbf{P} by $\bar{\mathbf{P}}_L$, the P -values are independent of the noise PSD. An accurate computation of the P -values is made possible through the standardization of \mathbf{P} by $\bar{\mathbf{P}}_L$ and computation of the false alarm rates for HC and BJ can be computed as in [Moscovich et al., 2016].

3.4 Standardization using parametric estimates of S_E

Estimates of \hat{S}_E different from the averaged periodogram \bar{P}_L can be used for standardization in (2.75). Sec. 2.6 has reviewed some methods to obtain such estimates. For the purpose of comparing standardization $\tilde{\mathbf{P}}|\hat{S}_E$ with $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$, we opt for AR parametric estimates allowing for an automatic parameter setting, as this approach is commonly used in practice.

3.4.1 AR parameters estimation

When using an estimate based on an AR process of order o , this order can be estimated based on statistical considerations using many criteria, *e.g.* [Akaike, 1969, Akaike, 1974, Parzen, 1975, Rissanen, 1984, Ding et al., 2015], which are defined respectively by:

$$\begin{aligned}
\text{Akaike's final prediction error:} & \quad FPE(o) := \hat{\sigma}_o^2 \left(\frac{N+o+1}{N-p-1} \right); \\
\text{Akaike's information criterion:} & \quad AIC(o) := \ln(\hat{\sigma}_o^2) + \frac{2(o+1)}{N}; \\
\text{Parzen's criterion AR transfer function:} & \quad CAT(o) := \left(\frac{1}{N} \sum_{j=1}^o \frac{N-j}{N\hat{\sigma}_j^2} \right) - \frac{1}{\hat{\sigma}_j^2}; \\
\text{Rissanen's minimum description length:} & \quad MDL(o) := \hat{\sigma}_o^2 \left(1 + \frac{o+1}{N} \ln(N) \right); \\
\text{Ding's Bridge criterion:} & \quad G(o) := \log \hat{\sigma}_o^2 + 2 \frac{o}{N} \sum_{i=1}^o \frac{1}{i}
\end{aligned} \tag{3.43}$$

with $o = 1, \dots, o_{max}$ the model orders with maximum value o_{max} , which has to be chosen, and $\hat{\sigma}_o^2$ the estimate of the prediction error power [Akaike, 1969]:

$$\hat{\sigma}_o^2 = \frac{1}{N} \sum_{t=o+1}^N \left(X(t) - \sum_{j=1}^o c_j X(t-j) \right)^2.$$

For each criterion, the order minimising the corresponding cost function in (3.43) is selected. In the numerical studies, we found that the selected order is often different from the true one (as expected, see e.g., p. 211 of [Akaike, 1969] and [Boardman et al., 2002] for similar conclusions). This is illustrated in the example below. However, as far as detection results are concerned, these criteria have very similar behaviour for sufficiently large N (cf. Sec. 3.6).

Example 13. Comparison between the different AR order estimation criteria

Fig. 3.4 shows histograms of the estimated order for the five criteria introduced in (3.43), evaluated with 10^4 MC simulations of AR processes with order $o = 6$ (blue histograms) and $o = 20$ (red histograms). The time series length is $N = 1024$ and the maximum order is set to $o_{max} = 30$. One can see that the FPE, AIC, CAT and G criteria often find selected orders different from the true ones. The CAT criterion overestimates the true order and G seems to detect with same probability all the considered orders values. In contrast, the MDL criterion most often found the true order for low value ($o = 6$) but fails to detect the larger order ($o = 20$).

Once the order has been estimated, the AR(o) coefficients are calculated by inverting the Yule-Walker equation (as this equation links the autocovariance matrix estimated from the data and the AR coefficients). Denoting by $\hat{\boldsymbol{\theta}}_{AR} := [\hat{o}, \hat{c}_j, \hat{\sigma}_o^2]^\top$ the AR parameter vector with respectively \hat{o}_{AR} the selected order, $\{\hat{c}_j\}_{j=1, \dots, \hat{o}_{AR}}$ the AR coefficients, and $\hat{\sigma}^2(\hat{o}_{AR})$ the estimated prediction error variance, the PSD estimate is obtained similarly as in (2.47) by:

$$\hat{S}_{E,AR}(\nu; \hat{\boldsymbol{\theta}}_{AR}) := \frac{\hat{\sigma}_o^2}{\left| 1 + \sum_{j=1}^{\hat{o}} \hat{c}_j e^{-2\pi i j \nu} \right|^2}. \tag{3.44}$$

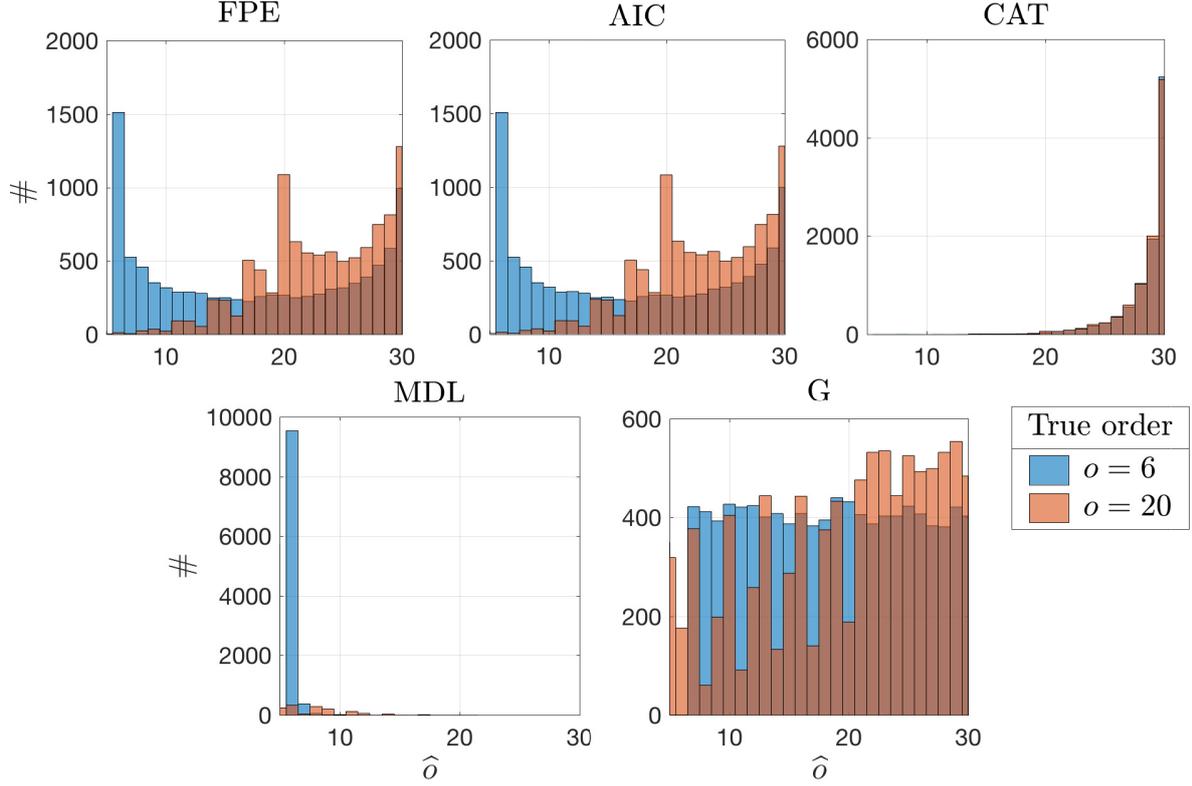


Figure 3.4 – Distribution of the estimated AR order estimates (\hat{o}) obtained with 10^4 MC simulations using a) FPE, b) AIC, c) CAT, d) MDL and e) G criterion. The blue histograms show the result for an AR(6) and the red for an AR(20). One can see that MDL criterion is better to find the true AR order for low o , while the other ones tend to overestimate its value.

3.4.2 Distribution of $\tilde{\mathbf{P}}|\hat{\mathbf{S}}_{E,AR}$

Even if such estimation approaches are relatively straightforward to implement, characterizing precisely the distribution of $\tilde{P}(\nu_k|\hat{\mathbf{S}}_{E,AR})$ is more difficult than in the case of $\tilde{P}(\nu_k|\bar{P}_L)$, owing essentially to the stochastic nature of \hat{o} in (3.44).

In practice however, the “whitening” effect of $\hat{\mathbf{S}}_{E,AR}$ is efficient because the selection procedures have good fitting properties, i.e. they are approximately consistent:

$$\hat{\mathbf{S}}_{E,AR} \xrightarrow{\approx} S_E \text{ as } N \rightarrow +\infty.$$

This leads to consider as reasonable the assumption that:

$$\frac{P(\nu_k)}{\hat{\mathbf{S}}_{E,AR}(\nu_k)} \approx \frac{P(\nu_k)}{S_E(\nu_k)} \quad (3.45)$$

and, with (3.4), that $\frac{P(\nu_k)}{\hat{\mathbf{S}}_{E,AR}(\nu_k)}$ is approximately a χ_2^2 r.v. for $k \in \Omega$ (and a χ_1^2 r.v. for $k = 0, \frac{N}{2}$):

$$\frac{P(\nu_k)}{S_E(\nu_k)} \sim \frac{\chi_2^2}{2} \quad \cong \quad \frac{P(\nu_k)}{\hat{\mathbf{S}}_{E,AR}(\nu_k)} \sim \frac{\chi_2^2}{2}. \quad (3.46)$$

This approximation, which may seem natural, can be used to evaluate false alarm rates. For example, using (3.46) and following the lines of (3.38), the false alarm of test statistics T_C

applied to $\frac{\mathbf{P}}{\widehat{\mathbf{S}}_{E,AR}}$ leads, under the above approximation, to:

$$\begin{aligned}
P_{FA}(T_C(\widetilde{\mathbf{P}}|\widehat{\mathbf{S}}_{E,AR}, N_C), \gamma) &:= \Pr(T_C(\widetilde{\mathbf{P}}|\widehat{\mathbf{S}}_{E,AR}, N_C) > \gamma | \mathcal{H}_0) \\
&\approx \Pr(T_C(\widetilde{\mathbf{P}}|\mathbf{S}_E, N_C) > \gamma | \mathcal{H}_0) \\
&= 1 - \sum_{k=0}^{N_C-1} \Pr(K = k | \mathcal{H}_0) \\
&= 1 - \Phi_{Bin}(u, \frac{N}{2} - N_C, N_C) \\
&= 1 - I_{1-u}(\frac{N}{2} - N_C, N_C) \\
&= I_u(N_C, \frac{N}{2} - N_C),
\end{aligned} \tag{3.47}$$

with $u := \Phi_{\chi^2_2}(2\gamma) = e^{-\gamma}$. The reliability of this approximation will be studied in Sec. 3.5.4. We will see that, even if the parametric model is well chosen, the resulting P_{FA} control is less reliable than the one derived by the non parametric standardization based on \overline{P}_L .

3.5 Numerical studies on synthetic noise processes

This section is a numerical study of the behavior of the tests statistics when applied to the proposed standardized periodogram $\widetilde{\mathbf{P}}|\overline{\mathbf{P}}_L$ (3.2) (cf. Sec. 3.3). We also cross compare the performance of various detection tests applied to $\widetilde{\mathbf{P}}|\overline{\mathbf{P}}_L$ and $\widetilde{\mathbf{P}}|\widehat{\mathbf{S}}_E$.

For the purpose of comparing theoretical results against a large number of simulations, a synthetic colored noise process is considered under \mathcal{H}_0 of model (3.1). As a common method to generate correlated noise, this frequency dependent noise is generated here as an AR process with order $o = 6$ and coefficients $\{c_j\}_{j=1,\dots,o}$ defined by:

$$\begin{aligned}
X(t) &= \sum_{j=1}^o c_j X_{t-j} + \epsilon(t) \\
&= 0.7X_{t-1} + 0.05X_{t-2} + 0.3X_{t-4} - 0.3X_{t-6} + \epsilon(t),
\end{aligned} \tag{3.48}$$

with $\epsilon(t) \sim \mathcal{N}(0, 1)$. In absence of any other specification, the time series' length is $N = 1024$ and $\Delta t = 1$ unit (for the purpose of making an illustration without consideration of any real physical timescale, one considers 1 unit is 1 second).

The coefficients of (3.48) have been chosen to yield a correlated noise process sketching the PSD of some stars, exhibiting higher energy at low frequencies (e.g. convection), and local variations at high frequencies (e.g. incoherent stellar oscillations). In the scope of the present study, this colored noise is intended to roughly represent the large frequency variations encountered in some stellar PSD. The choice of this PSD is obviously not intended to reflect the reality of a particular stellar type.

One realization of this noise process is shown in Fig. 3.5 (panel a). The associated PSD is shown in panel b) with the theoretical PSD, given in (2.47), overplotted in red. Panel c) shows the theoretical autocorrelation function r_E (evaluated as described in Appendix. C). One can observe the fast decay of the noise correlation as a function of the lag, and r_E totally vanishes around a lag of about 40 samples.

The theoretical evaluations of the previous sections call for various interesting comparisons. The following numerical study is divided in five parts.

First, in Sec. 3.5.1, the analytical expressions of the P_{FA} and P_{DET} of tests T_M and T_C applied to $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$ derived in Sec. 3.3.1 and Sec. 3.3 are verified with MC simulations. Then, we study the influence of different parameters related to the nature of the data (data length and noise autocorrelation function) and to the parameter setting (number of available noise training time series and estimated number of harmonics).

In Sec. 3.5.2, the performance of the adaptive tests HC (2.70) and BJ (2.74), applied to the standardized periodogram P -values (3.42), are studied for different sparsity regimes under the alternative and compared to T_M and T_C .

In Sec. 3.5.3 and Sec. 3.5.4, the proposed methodology based on the standardized periodogram (3.2) is compared with an approach ignoring the noise correlation, and with an approach assuming a parametric model on the noise process.

Finally, in Sec. 3.5.5 an application of this work to RV using HD simulated time series (as described in Sec. 1.4) will be described.

In this section, the units of the different quantities have to be considered without any physical realism.

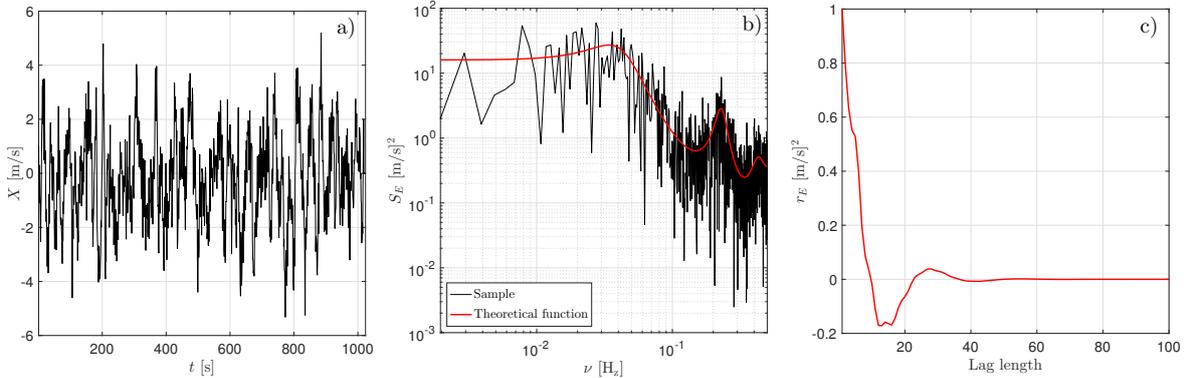


Figure 3.5 – Illustration of the considered AR(6) process through a) one realization of this process, b) power spectral density and c) autocorrelation function. In panel b) the red curve illustrates the theoretical PSD given in (2.47).

3.5.1 Classical tests based on $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$

We first consider the theoretical results obtained for tests $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ and $T_C(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ defined in (3.11) and (3.31).

False alarm and detection probabilities

Under the null of model (3.1), 10^4 MC simulations of the considered AR(6) noise are generated. Under the alternative, $N_S = 3$ sinusoids have been added with equal amplitudes $\alpha_s = 0.2 \text{ m}\cdot\text{s}^{-1}$, frequencies $f_s = 0.05, 0.12$ and 0.22 Hz and random phases in $[0, 2\pi]$. For each time series, the tests $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ and $T_C(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ (using $N_C = N_S = 3$) are applied. The empirical probabilities are computed by:

$$P_{FA}(\gamma) = 1 - \widehat{\Phi}_T(\gamma|\mathcal{H}_0)$$

$$P_{DET}(\gamma) = 1 - \widehat{\Phi}_T(\gamma|\mathcal{H}_1)$$

with $\widehat{\Phi}_T$ the empirical cdf (obtained with 10^4 realizations) of the considered test statistics under both hypothesis.

Fig. 3.6 aims to evaluate the accuracy of the analytical expressions of the false alarm given in (3.18) and (3.38) (panel a)), and detection probabilities given in (3.19) and (3.41) (panel b)) derived for both tests. The ROC curves are shown in panel c).

All the theoretical expressions are shown by the colored dot lines and the empirical results by the associated colored solid lines.

For all the considered parameters, one observes a fair agreement between the empirical and theoretical (derived for the asymptotic regime) results, even for the not so large value of N considered here ($N = 1024$).

Different values of L are shown here to illustrate the improvement brought by larger training data sets ($L = 5$ for the blue and green curves, $L = 100$ for the black and red curves). One can observe that both tests performance logically increase with L , as the estimation noise decreases with the increasing size of the training data set.

Regarding the ROC (panel c)), the test T_M outperforms T_C for all considered values of L . This behavior would be different if N_S would be larger (cf. Fig. 3.7).

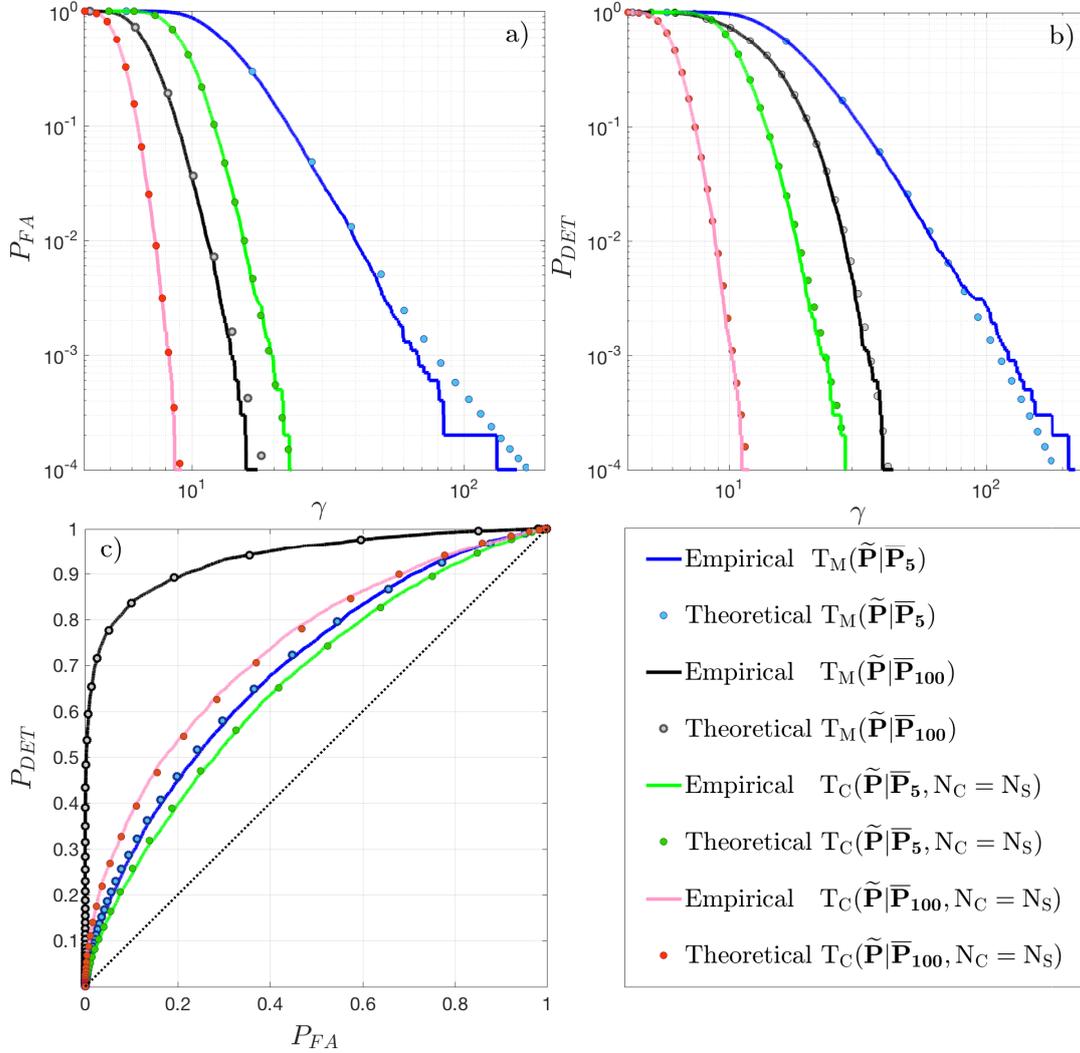


Figure 3.6 – Theoretical (solid lines) *vs* empirical (dot lines) results for $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ and $T_C(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$: a) FA probability (expressions (3.18) and (3.38)), b) detection probabilities (expressions (3.19) and (3.41)), and c) ROC curves (see (3.21)).

We compare now the performance of T_M and T_C depending on the parameters L and N_C .

Influence of L and N_C

The performance of tests $T_M(\tilde{\mathbf{P}}|\overline{\mathbf{P}}_L)$ and $T_C(\tilde{\mathbf{P}}|\overline{\mathbf{P}}_L)$ depend on the number of available training noise time series (L) and on the number of sinusoids presumably contaminated by noise (N_C). An interesting question is the gain brought by T_C over T_M in case of multiple sinusoids truly present under \mathcal{H}_1 (note that no gain is apparent for $N_S = 3$ in Fig. 3.6, panel c)). Under \mathcal{H}_1 , we consider $N_S = 20$ sinusoidal signals (in the same AR(6) noise) with equal amplitudes $\alpha_s = 0.05 \text{ m.s}^{-1}$ and frequencies $\{f_j\}_{j=1,\dots,N_S}$ equally spaced in the interval $[4.5, 8]$ mHz (not fixed on the Fourier grid). The results are shown in Fig. 3.7.

Consider first the influence of L . The panel a) shows the empirical ROC curves obtained for tests $T_M(\tilde{\mathbf{P}}|\overline{\mathbf{P}}_L)$ (black) and $T_C(\tilde{\mathbf{P}}|\overline{\mathbf{P}}_L)$ (red). The standardized Fisher's test T_F applied to $\tilde{\mathbf{P}}|\overline{\mathbf{P}}_L$ (yellow) is also shown (only for comparison as computing the corresponding P_{FA} is not possible, cf. Sec. 3.3.2). The different values of L (5, 20, 100) are indicated by different linestyles (see legend). The N_C parameter of T_C is set to N_S . One observes that, as expected, test T_C is more powerful than T_M when multiple sinusoids are present under \mathcal{H}_1 . This will be true if N_S is significantly larger than 1, but this also depends on N_C parameter. The question of the influence of N_C for this test is thus now investigated.

The panel b) of Fig. 3.7 shows the empirical ROC curves of $T_C(\tilde{\mathbf{P}}|\overline{\mathbf{P}}_L)$ for the same signal (i.e. $N_S = 20$ sines in the AR(6) noise), with $L = 100$ and $N_C = 1, 10, 20 (= N_S), 30$ and 80 (see legend). These numerical results show that the ROC curves are stable over a relatively wide range of N_C values around N_S . This test is thus not very sensitive to the N_C parameter. The best performance of T_C are obtained for values of N_C close to the real number of sinusoids. Note that the value $N_C = N_S$ does not yield the best performance. This shows that the region to look at in order to “make the case” against the null may not exactly be the N_S^{th} largest peak (or equivalently, the N_S^{th} smallest P -value), but slightly further away. This somewhat surprising fact was noticed in [Donoho and Jin, 2004].

When a rough guess of the signal under \mathcal{H}_1 is available, it may be possible to adjust the value of N_C without MC simulations, by comparing ROC curves obtained from (3.38) and (3.41). However, in cases where no relevant guess about N_S can be made *a priori*, it may be useful to turn to adaptive approaches (see Sec. 3.5.4). Adaptive *vs* non adaptive approaches (i.e. T_C *vs* HC and BJ) will be compared in Sec. 3.5.2.

To conclude, the tests T_M and T_C applied to $\tilde{\mathbf{P}}|\overline{\mathbf{P}}_L$ allow to control precisely the P_{FA} even for a finite number of data for the considered type of noise correlation timescales. Their performance logically increase with L . The performance of test T_C are not very sensitive to the setting of parameter N_C and this test outperforms T_M as long as the number of harmonic oscillations under \mathcal{H}_1 is sufficiently large w.r.t. 1 ($N_S = 20$ in our example).

Influence of the length of the time series (N)

Regarding now the influence of N (as our results are derived in the asymptotic regime), we compare the theoretical probabilities of test $T_M(\tilde{\mathbf{P}}|\overline{\mathbf{P}}_L)$ given in (3.18) and (3.19) with empirical probabilities obtained for time series of three different lengths N . We consider the AR(6) process under the null and a single sinusoid on the Fourier grid under the alternative. For the three time series of different lengths the sinusoid amplitude is resp. $\alpha_s = 0.4, 0.25$ and 0.1 m.s^{-1} , the frequency is a Fourier frequency $f_s = \nu_{370} = 0.1 \text{ Hz}$, the phase is null, and $L = 20$ training data sets have been used. On these simulations, the parameter α_s differs only for the ROC curve visibility and does not influence the following results.

Fig. 3.8 shows the ROC curves of test $T_M(\tilde{\mathbf{P}}|\overline{\mathbf{P}}_L)$ for three values of $N = 100, 1024$ and 2500 (resp. from left to right). The red curves indicate the theoretical values (given in (3.22), (3.38))

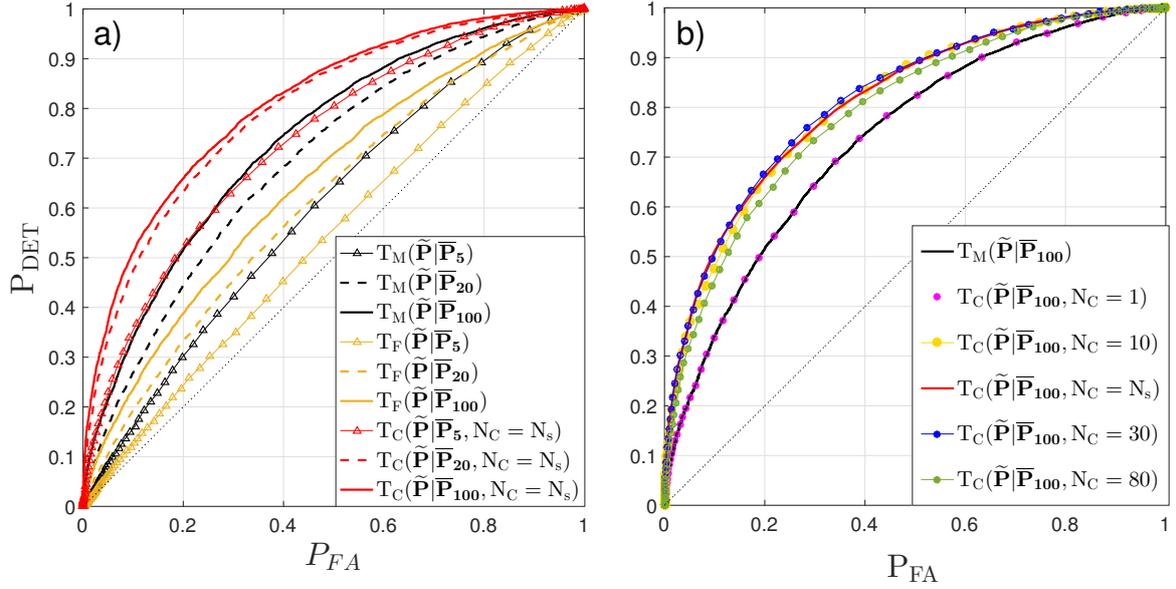


Figure 3.7 – ROC curves of tests $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ (3.11) and $T_C(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ (3.31) showing the influence of a) the number of available light curves (L) and b) the parameter N_C on the tests performance.

and (3.41)) while the empirical values obtained through 10^4 MC simulations are shown in black.

Fig. 3.9 shows the absolute differences between theory and empirical estimates as a function of the threshold γ for the P_{FA} (left) and P_{DET} (right) with the different colors comparing the deviations for the different N values. One can observe from these two figures the decrease of the maximal error as N increases, as expected as our results are asymptotic. These deviations have however the same order of magnitude (about 5×10^{-3} for all γ) and remain not significant even for the case of a low number of observations. To conclude, the theoretical probabilities derived in the asymptotic regime are also relevant for finite N values. Of course, this depends on the correlations timescales w.r.t. N and we now investigate this point.

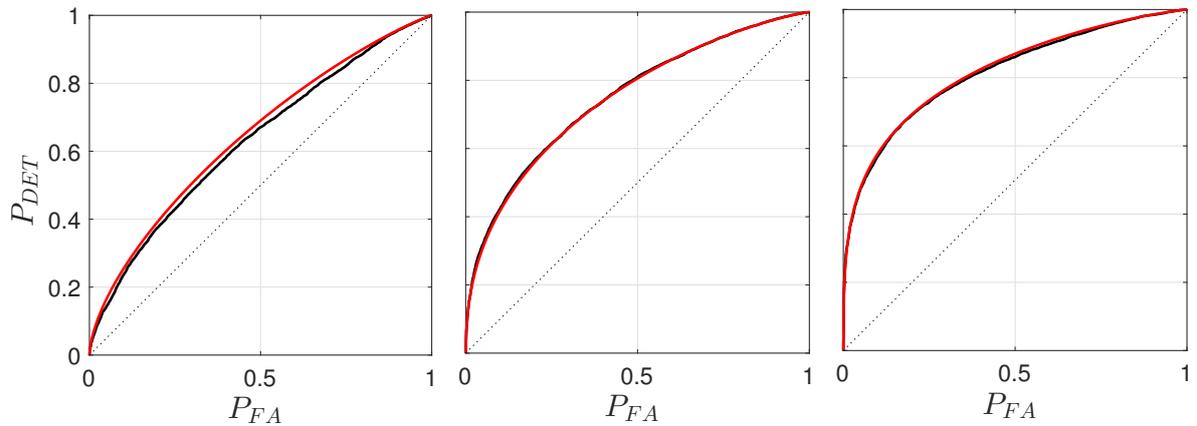


Figure 3.8 – Theoretical (red) *vs* empirical (black) ROC curves of $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ (3.11) for $N = 100, 1024$ and 2500 (resp. from left to right). One can observe the decrease of the ROC deviations for large N .

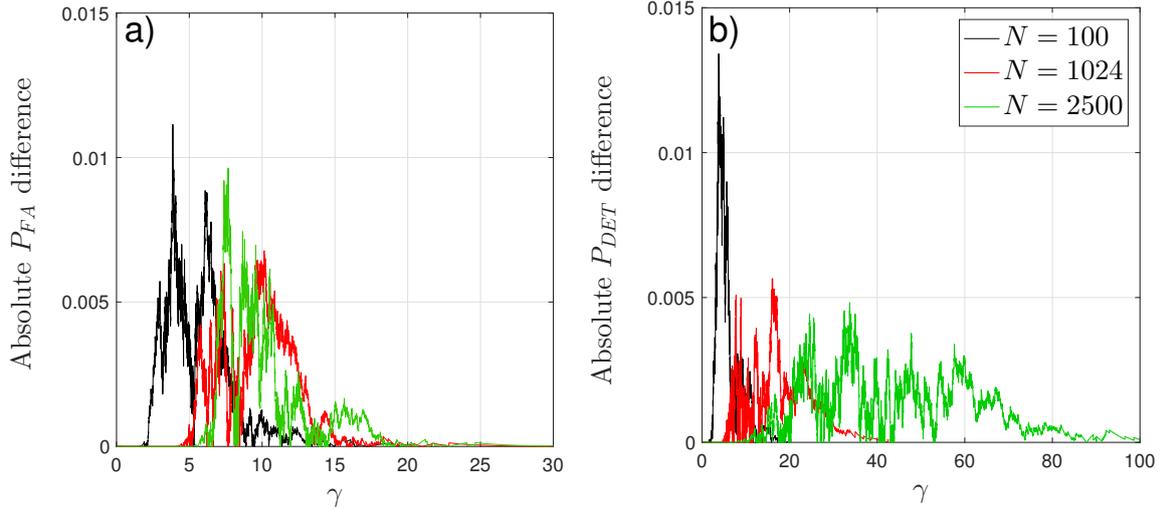


Figure 3.9 – Differences between theory and empirical evaluation of the P_{FA} (left) and P_{DET} (right) in Fig. 3.8 as a function of γ . One can see the decrease of both quantities with large N .

Influence of the noise autocorrelation function: bias and dependencies

We investigate now more precisely the influence of several factors on the theoretical expressions derived for test $T_M(\tilde{\mathbf{P}}|\overline{\mathbf{P}}_L)$ in (3.18) and (3.19). The factors explaining possible deviations between theory and empirical results are:

- large correlations (through r_E) leading to $\rho_k \neq 1$ (cf. Sec. 2.2.6, Theorem. 1),
- the existence of a small bias at finite N (cf. Sec. 2.2.5),
- dependencies of the ordinates $\tilde{\mathbf{P}}|\overline{\mathbf{P}}_L$ (cf. Sec. 2.2.6 and Theorems 5.2.1 and 5.2.4 of [Brillinger, 1981] on the asymptotic independence between the periodogram ordinates).

Under \mathcal{H}_1 , consider a single sinusoid ($\alpha_s = 0.25 \text{ m.s}^{-1}$, $f_s = \nu_{370} = 0.1 \text{ Hz}$, $\varphi_s = 0$) in three different order AR processes:

- the AR(6) process, introduced in (3.48), with $N = 1024$ (green curves in panels b) and c) in Fig. 3.10),
- an AR(100) process with $N = 1024$ (red curves in panels b) and c) in Fig. 3.10),
- an AR(100) process with $N = 500$ (black curves in panels b) and c) in Fig. 3.10).

Panel a) of Fig. 3.10 shows the theoretical autocorrelation functions of the considered processes ($r_E(o = 6)$ in red and $r_E(o = 100)$ in black). The theoretical AR autocorrelation functions r_E are computed using the algorithm described in Appendix. C. One can see the correlations of the high order AR process extend until a lag of ≈ 150 samples and the AR(6) until a lag of ≈ 40 (as also seen in Fig. 3.5, panel c)).

Panels b) and c) shows respectively the absolute difference between the empirical and theoretical P_{FA} and P_{DET} as a function of γ for $T_M(\tilde{\mathbf{P}}|\overline{\mathbf{P}}_L)$. The theoretical results have been obtained through 10^4 MC simulation of the three different AR processes. In panel c), the solid curves illustrate the case where the theoretical P_{DET} have been computed with (solid lines) and without (dashed lines) the approximation $\rho_k(r_E) \approx 1$ in (3.19) and (3.6). In the case where ρ_k is not approximated, the term has been evaluated as (cf. Appendix. B and Theorem 6.4, p.178 of [Li, 2014]):

$$\rho_k := \frac{S_E(\nu_k)}{N\sigma_k^2}, \quad (3.49)$$

with σ_k^2 given in (B.5) as

$$\sigma_k^2 := \frac{1}{N} \sum_{t,s=1}^N r_E(t-s) e^{-i2\pi\nu_k(t-s)}. \quad (3.50)$$

One can observe a larger mismatch between theory and observation under both hypotheses for N low (black *vs* green/red).

Under \mathcal{H}_1 in panel c), one sees that the P_{DET} differences have globally the same order of magnitude as the P_{FA} differences. Small differences (less than 1%) appear for the highly correlated case (AR(100)). There is no improvement when considering the correct ρ_k (solid *vs* dashed) suggesting that the mains sources of mismatch at finite N are the small dependencies between the periodogram ordinates in (3.5) and the bias (the marginal distributions are not exactly $S_E\chi_2^2/2$ at finite N).

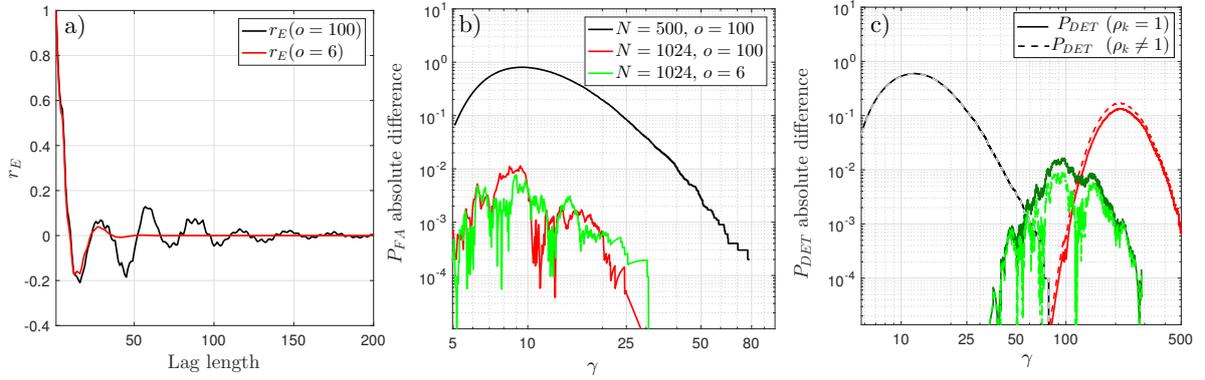


Figure 3.10 – Illustration of different order of correlations on the theoretical results. a) Autocorrelation function of the considered AR(6) (red) and AR(100) (black). b) P_{FA} absolute differences between empirical and theoretical (3.18) expressions depending on the noise correlation (through o) and N . c) P_{DET} differences with the theoretical expression (3.19) evaluated with (solid lines) and without (dashed lines) the approximation $\rho_k \approx 1$.

To conclude, the non vanishing bias and dependencies at finite N are the main limiting factors to the reliability of the theoretical expressions derived in Sec. 3.2. However, the mismatch between theory and practice is very small for correlation timescale smaller than N (cf. Fig. 3.6 and Fig. 3.9) and the theoretical results reflect very closely the performance observed in practice. Regarding the considered application, where the colored noise is due to the convective phenomena, this noise is correlated on timescales of minutes to days while the observations are sampled each 1 or 3 points by night over several months in general. The theoretical results above are then reliable in this regime (for a regular sampling).

3.5.2 Adaptive tests based on $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$

When multiple deviations occur under the alternative, the proposed standardization allows the use adaptive tests, as HC (2.70) and BJ (2.74), which are designed to detect non extremely sparse signatures.

Compared ROC curves

Considering such adaptive tests applied to standardized data $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$ (3.2), one first wants to study their performance and compare them to those tests T_M and T_C for different signal sparsity regimes in the Fourier space.

Fig. 3.11 shows the comparison results for different Keplerian planetary signatures (computed using (1.3)) under \mathcal{H}_1 and the AR(6) noise described in (3.48) under \mathcal{H}_0 .

Consider first the case where a single sparse signal is present under the alternative, i.e. when a single planet signature with zero eccentricity is present in the periodogram. The periodogram of the sparse sinusoidal signal is shown in panel c). Different signal RV semi-amplitudes have been considered for the ROC visibility (we plot in red the signal used for $L = 1$ and in black the one used for $L = 50$). The used Keplerian parameters are $M_p = [10 M_\oplus, 2.5 M_\oplus]^\top$ (resp. for $L = 1$, dashed curves, and $L = 50$, solid curves in a)), $T_p = 5.33$ days ($f_p = 2.17 \mu\text{Hz}$ on grid, $N = 1800$, $\Delta t = 30$ min). The other Keplerian parameters are null. Panel a) shows the respective empirical ROC curves of T_M (black), T_C (red), HC (blue) and BJ (green). In the adaptive tests, half of the independent periodogram P -values have been considered ($\alpha_0 = 1/2$, i.e. $k = 1, \dots, \frac{N/2-1}{2}$). Note that when $N_C = N_S = 1$, T_C reduces to T_M (cf. (2.61)). In this case, one can observe that T_M is slightly more powerful than the adaptive tests.

This situation changes when the number of harmonics is large (though still small w.r.t. N), which is expected (cf. Sec. 3.5.1). This is shown in panel b) where $N_p = 5$ RV planetary signals have been introduced. The used Keplerian parameters for the five planets are respectively, for $p = 1, \dots, N_p$: $M_p = [0.5 M_\oplus, 0.5 M_\oplus, 1 M_\oplus, 1 M_\oplus, 1 M_\oplus]^\top$ (for $L = 1$, dashed curves in b)) and $M_p = [0.5 M_\oplus, 0.5 M_\oplus, 0.7 M_\oplus, 0.8 M_\oplus, 0.8 M_\oplus]^\top$ (for $L = 50$, solid curves), $T_p = [0.46, 1.3, 2.4, 6.9, 9.2]^\top$ days (f_p off-grid, such $f_p = [24.77, 8.69, 4.73, 1.67, 1.25]^\top \mu\text{Hz}$), $e_p = 0.9 \forall p$, and $\omega_p = \pi \forall p$ (the other parameters are null). The periodogram of the pure Keplerian signals is shown in d). One can visualize several harmonics appearing in the periodogram due both to the large eccentricities and off-grid planet frequencies (here, we do use any criterion to count the number of harmonics as, strictly speaking, it is equal to N , but one can observe in this example more than 50 harmonics > 1 [m/s] 2 for $L = 1$, and more than 20 harmonics > 1 [m/s] 2 for $L = 50$). An interesting point is the comparison of HC and BJ with $T_C(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L, N_C)$. For this latter test, N_C is set to N_p . Hence, T_C represents a kind of Oracle, which knows in which region of the P -values to look at in order to “make the case” against \mathcal{H}_0 (for this test, best performance are obtained for N_C close to N_p , cf. Sec. 3.5.1). Panel b) shows that adaptive procedures (and in particular BJ) have a power higher than T_C , for all considered values of L , even if they do not benefit from such *a priori* knowledge. This makes such tests very interesting in the case of unknown sparsity in the Fourier space.

Probability of detecting the true support in the Fourier space

While T_M is, by construction, focused on the first ordered P -value (or equivalently, on the largest peak in the standardized periodogram), adaptive tests allow to detect other deviations from \mathcal{H}_0 corresponding to several peaks of the standardized periodogram. As we have seen in Fig. 3.11, adaptive tests are efficient for detecting non extremely sparse signals in terms of ROC, we may ask what is their probability to correctly determine the location of the signal components in the Fourier space? The following study examines this question for model (3.1).

- **Study under the null.** We perform 10^4 MC simulations of the AR(6) process defined in (3.48), and apply tests $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_{20})$, $\text{HC}(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_{20})$ and $\text{BJ}(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_{20})$.

Our goal is to evaluate if some frequencies are preferentially identified by the three test under the null. For test T_M , we make the histogram of the frequencies where the maximum peak is found (left panel, top row of Fig. 3.12). For HC and BJ, these test look at the maximum of test statistics (HC_k and BJ_k in (2.70) and (2.74)). Let k_\star be the index achieving the maximal test statistics. This index corresponds to the frequency, say ν_{k_\star} the k_\star -th largest peak of the standardized periodogram, and we show the histograms of the collected frequencies ν_{k_\star} for both tests in the middle and right panels. One can observe a uniform distribution of all these tests over the frequencies. This desirable effect comes from the periodogram standardization in (3.2). This means that the tests are not

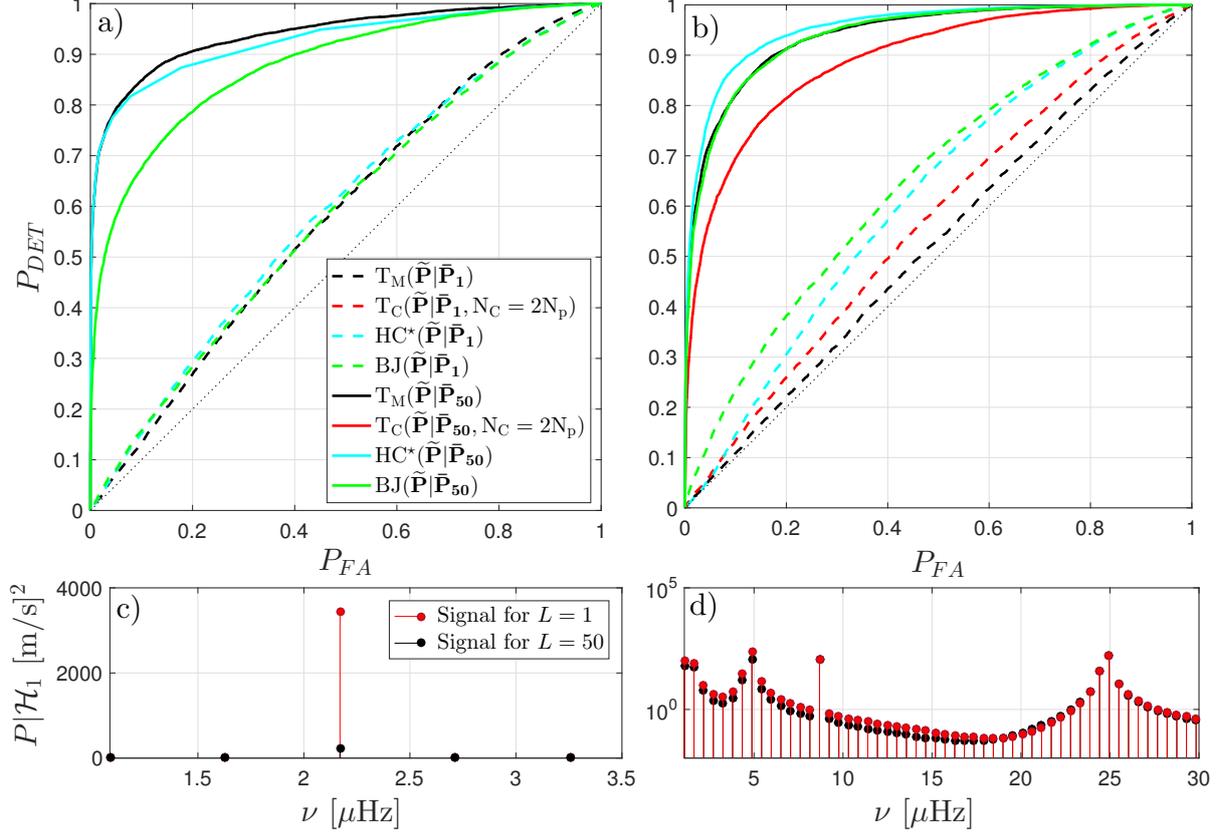


Figure 3.11 – Comparison of adaptive tests HC (2.70) and BJ (2.74) with classical tests T_M (3.11) and T_C (3.31) applied to standardized data $\tilde{\mathbf{P}}|\tilde{\mathbf{P}}_L$ for $L = 1$ (dashed ROC) and $L = 50$ (solid ROC). Panels a) and c) show the case of a sparse Keplerian signal ($N_p = 1$, null eccentricity, on grid planet frequency) under the alternative, computed thanks to (1.3). Panels b) and d) show the case of non extremely sparse signal ($N_p = 5$, large eccentricity, off grid planets frequency). The first row shows the ROC curves and the bottom row the periodogram of the pure sinusoids (without the AR noise).

biased toward any particular frequencies, even if the noise is colored.

The bottom row shows the resulting empirical $P_{FA}(\gamma)$ curves for the three tests (black). The theoretical P_{FA} expression derived for $T_M(\tilde{\mathbf{P}}|\tilde{\mathbf{P}}_L)$ in (3.18) is overplotted in dashed green. The red horizontal line indicates a P_{FA} of 5% obtained resp. at $\gamma_0 = 11.6$, 4.7 and 0.997 for resp. T_M , HC and BJ. One can observe that the three tests have very different distribution under \mathcal{H}_0 .

In practice, the FA threshold γ_0 at a given P_{FA} can be obtained theoretically through (3.21) for T_M and using MC simulations for the adaptive tests for which no analytical P_{FA} expression is obtainable. One can refer, for example, to Moscovitch’s work, who proposed a fast algorithm to compute P -values for such adaptive tests [Moscovitch et al., 2016].

- **Study under the alternative.** We recall that one common measure of sparsity under \mathcal{H}_1 is the proportion of nonzero coefficients parametrized as $N_S = N^{1-\beta}$, with $\beta \in [\frac{1}{2}, 1]$ the sparsity coefficient (cf. [Donoho and Jin, 2004] and Sec. 2.5.1). We follow here closely this parametrization, with N_S signals of equal amplitudes in the time domain (α) and frequencies f_j ($j = 1, \dots, N_S$) randomly chosen in the Fourier grid. Let $\Omega_S \subset \Omega$ define the set of N_S signal frequencies.

To evaluate the tests’ ability to locate the correct frequency support Ω_S , we generate

$N_{MC} = 10^4$ MC simulations ($i = 1, \dots, 10^4$) for varying sines' amplitudes and sparsity coefficient (β), so varying N_S . Locating the correct support means that the considered test, say T , leads to N_S ordinates found significant (which we note $\widehat{N}_S = N_S$) and the significant frequencies are all in Ω_S (which we note $\widehat{f}_j = f_j, \forall j = 1, \dots, N_S$) for a given threshold $\gamma = \gamma_0$ corresponding to a target P_{FA} .

For test T_M , the number of estimated components is indeed always $\widehat{N}_S = 1$ as it focuses only on the maximum periodogram value and the \widehat{f}_j is the frequency at which the periodogram is maximum. For adaptive tests, as seen in Example. 12 of Sec. 2.5.3, \widehat{N}_S can be measured as the number of ordered P -values smaller than the one selected by the tests (in (2.70) and (2.74)). The frequencies \widehat{f}_j correspond then to the periodogram frequencies associated to these \widehat{N}_S smallest ordered P -values (or, equivalently, to the \widehat{N}_S frequencies where the N_S largest peaks occur in the standardized periodogram).

The probability of detecting the correct support at the right location is evaluated through the 10^4 MC simulations for which one retains the number ($\#$) of times where the test T fits the three following conditions

$$\widehat{\Pr} \left(\{T > \gamma_0\} \cap \{\widehat{N}_S = N_S\} \cap \{\widehat{f}_j = f_j, \forall j\} \right) := \frac{\# \{T > \gamma_0 \ \& \ \widehat{N}_S = N_S \ \& \ \widehat{f}_j = f_j, \forall j\}}{N_{MC}}. \quad (3.51)$$

Fig. 3.13 represents this estimated probability as a function of the sparsity parameter $\beta(N_S) := 1 - \frac{\log N_S}{\log N}$ (2.66) and the sines' amplitudes. In this example, the number of harmonics is precisely countable as the signals are sinusoidal functions with on grid frequencies. For these simulations, we used $L = 20$ and the threshold of each test (T_M , HC, BJ, resp. from left to right) corresponds to a P_{FA} of 5% found empirically for HC and BJ thanks to MC under \mathcal{H}_0 (see Fig. 3.10, last rows, red lines).

A grid of parameters of size 500×500 is evaluated for signal amplitudes $\alpha \in [0.05, 10]$ m.s⁻¹ and $N_S \in [1, 100]$ (corresponding to $\beta \in [0.33, 1]$). For T_M , a zoom around $\beta = 1$ (i.e. $N_S = 1$) has been done for visibility. Indeed, this test identifies one frequency and has then a null probability to find the correct support if $N_S > 1$.

BJ has the best performance over the whole considered sparsity regime (i.e. on $0.33 \leq \beta \leq 1$). This test can locate the correct support with a high probability for signals with $\alpha > 6$ m.s⁻¹ and $\beta \in [0.33, 1]$. If the performance of HC and BJ could be close in terms of ROC (see Fig. 3.9) in the case of multiple signals, BJ better detects the true components at their true location in a large part of the sparsity regime¹.

¹These results have not been published yet.

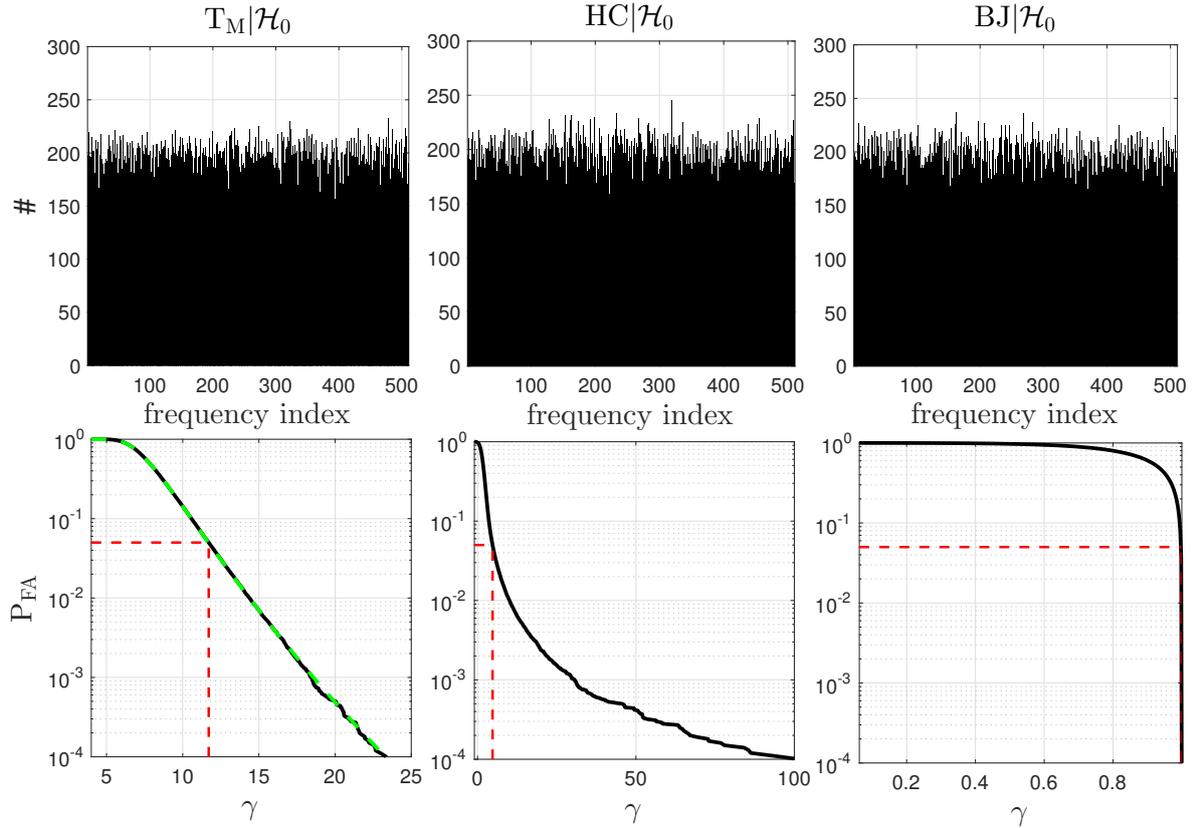


Figure 3.12 – Study under \mathcal{H}_0 of tests T_M (3.11), HC (2.70) and BJ (2.74) applied to $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_{20}$ (3.2). First row: “Frequency distribution” of each test statistics obtained through 10^4 MC simulations. Last row: Empirical $P_{FA}(\gamma)$. The dashed red curves indicates a P_{FA} of 5% obtained resp. at $\gamma_0 = 11.6, 4.7$ and 0.997 for the three considered tests.

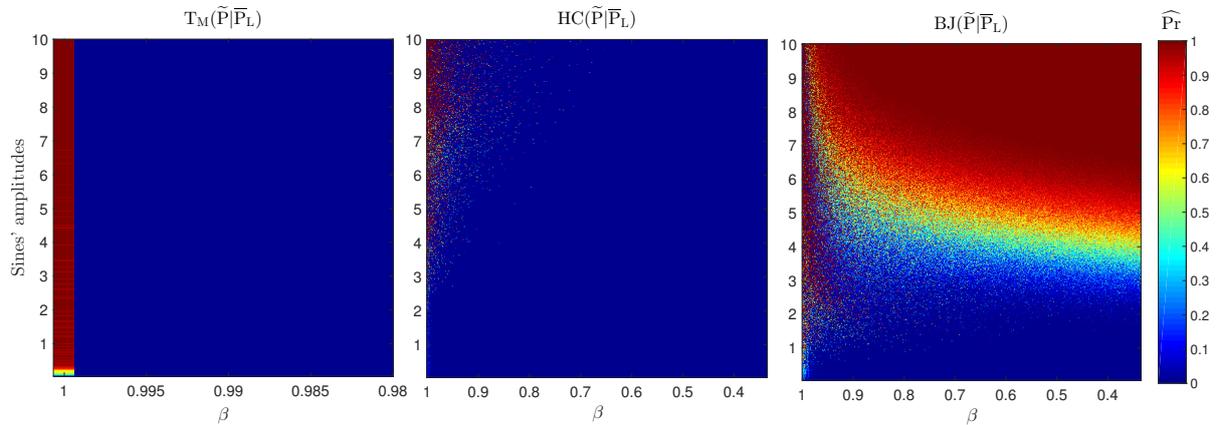


Figure 3.13 – Probability of detecting the correct support under \mathcal{H}_1 for T_M (left), HC (middle) and BJ (right).

3.5.3 Effect of ignoring noise correlations

In Chap.2 (Sec. 2.6), we showed how ignoring noise correlations and applying statistical tests to a classical (non standardized) periodogram leads to unreliable detection outcomes. Let us illustrate this effect on the colored AR(6) noise described in (3.48) ($\Delta t = 60$ s here).

Consider two experiments where the sinusoidal signals are located in two opposite regions of the noise PSD of the AR(6) process (cf. Fig. 3.5, panel b). The results are shown in Fig. 3.14.

First experiment: Under the alternative, $N_S = 5$ sines are placed at signal frequencies in a “bump” of the noise PSD (red lines, panel a). Panel b) shows the resulting empirical ROC curves for tests T_M , T_F , T_C and T_{Ch} (cf. Sec. 2.4). One sees that the performance of classical tests like T_{Ch} (cyan) and T_F (purple) applied to \mathbf{P} appear more powerful than T_M (black), T_F (yellow) and T_C (red) applied to $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$.

Second experiment: Under the alternative, the $N_S = 5$ signal frequencies fall in a “valley” of the noise PSD. The empirical ROC curves are shown in panel c). One sees that the performance of tests using $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$ are now better than those based on \mathbf{P} in this configuration (the classical test ROC curves are on the diagonal).

To gain more insight on the relative tests performance, one compares, in the first row of Fig. 3.15, the frequency distribution under \mathcal{H}_0 (as for Fig. 3.12, top row) of tests $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$, $T_C(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$, $T_F(\mathbf{P})$, and $T_{Ch}(\mathbf{P})$ (from left to right). For T_F and T_{Ch} , the FA repartition in frequency is not uniform and increases in the PSD regions of larger energy ($\nu < 1$ mHz). When signals frequencies happen to fall into these zones, tests based on \mathbf{P} are favoured, but when the frequencies fall outside such regions they have vanishing power. In contrast, the T_M and T_C tests applied to $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$ allow a good detectability over all the frequency range, with performance close to the asymptotic ones ($L \rightarrow \infty$, no estimation noise) for $L \approx 100$. Under \mathcal{H}_1 , one can see that all the tests focus on the noise PSD bump when the signals are in these regions (second row). When the signals fall into the valley regions (last row), tests based on P fail to detect the signals.

The false alarms of “unstandardized tests” accumulate precisely in the regions where the noise PSD is large. Conversely, such tests have vanishing detection power in the PSD “valleys”. In contrast, because of standardization by $\bar{\mathbf{P}}_L$, the tests $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ and $T_C(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ have constant false alarm rate over the whole frequency band (as also seen for HC and BJ applied to $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$ in Fig. 3.12).

As a final remark, note that the tests based on \mathbf{P} cannot be implemented as no threshold γ , corresponding to a target P_{FA} , can be derived. This is because the ROC curves can only be derived if the P_{FA} and P_{DET} of the considered test are known, which is not the case for tests based on P when the noise is colored.

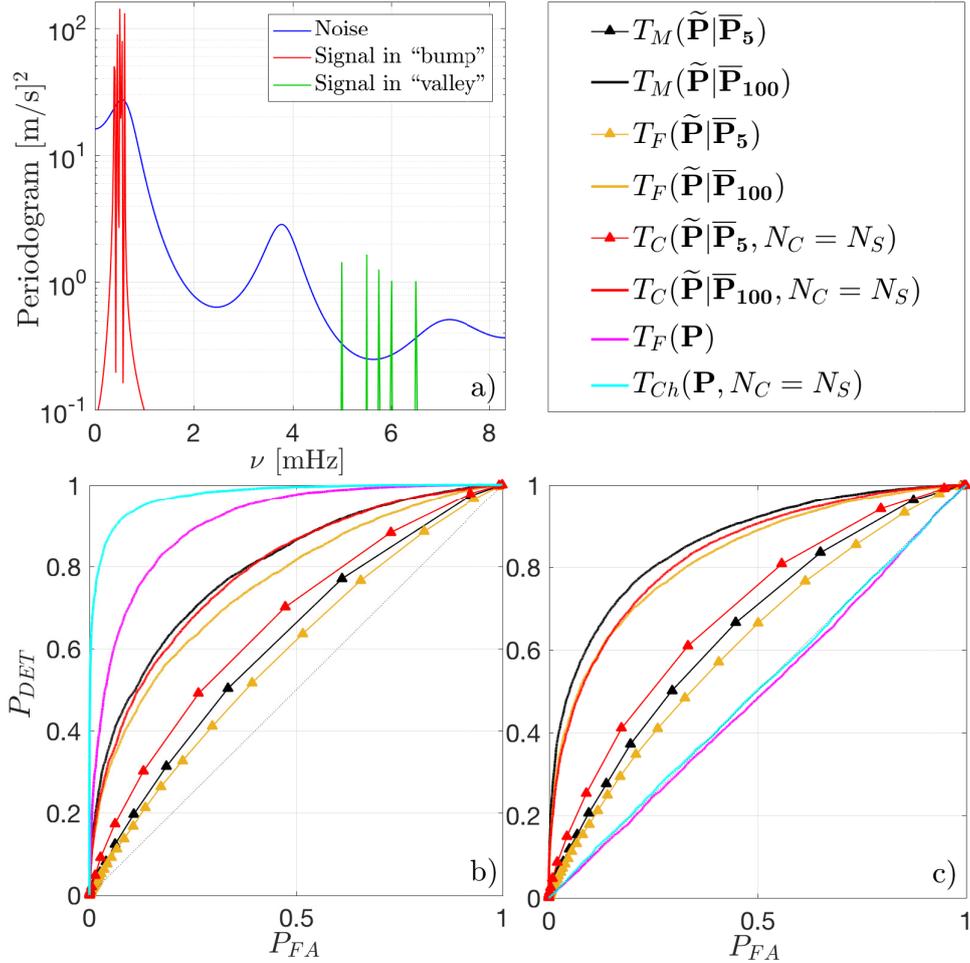


Figure 3.14 – (a) Noise and signals PSD. (b-c) ROC curves in case of signal frequencies falling into the PSD “bump” region or the “valley” region of the noise PSD. The sines amplitudes have been changed in the two cases for the ROC visibility but that does not influence the results.

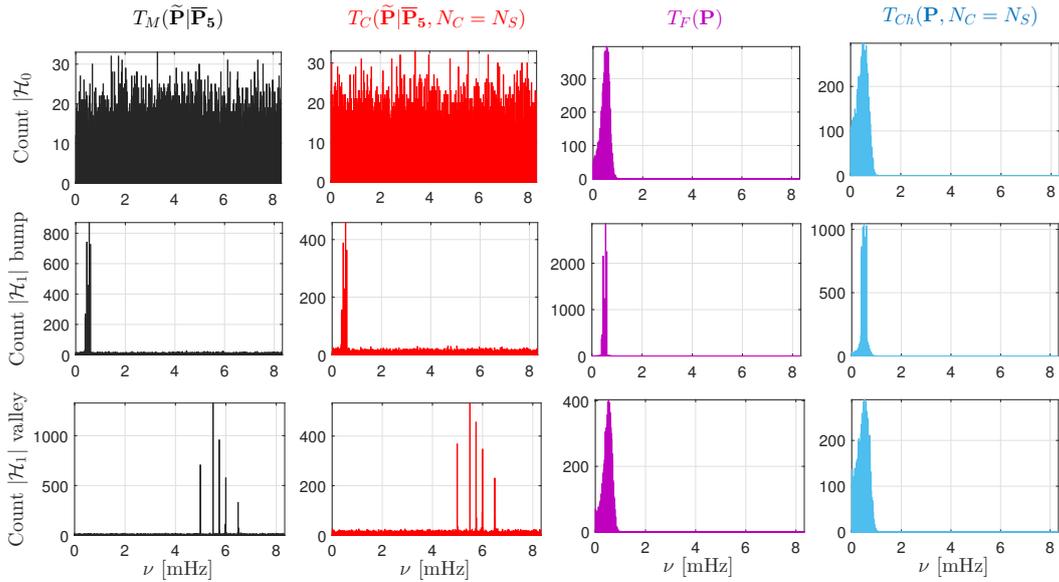


Figure 3.15 – First row: Histograms of the frequency distribution under \mathcal{H}_0 of (from left to right) tests $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$, $T_C(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$, $T_F(\mathbf{P})$ and $T_{Ch}(\mathbf{P})$. Last rows: Distribution under \mathcal{H}_1 for the case where the signals frequencies fall into a noise “bump” region or a “valley” region.

3.5.4 Comparison of the proposed standardization with a parametric approach

As discussed in Sec. 3.4, a way to deal with the frequency dependence of the noise is to estimate its PSD by parametric models.

In order to compare performance of standardization by $\bar{\mathbf{P}}_L$ and by a parametric estimate independently from the choice of the noise model, we will apply estimation methods specifically dedicated to AR processes (i.e. the chosen model is the true one and the parameters are estimated as in (3.43)).

The question is the reliability of tests using $\tilde{\mathbf{P}}|\hat{\mathbf{S}}_{E,AR}$, and in particular how accurate is approximation (3.47) for the P_{FA} with this approach. (Recall that a difficulty with such methods is the injection of estimation noise in the detection process through the parameters involved in $\hat{\mathbf{S}}_{E,AR}$. A standard approach discussed in Sec. 2.6 is to consider that the estimates are sufficiently accurate for their intrinsic variance and bias to be negligible, i.e. $S_E \approx \hat{S}_{E,AR}$. We study the performance of this approach here.) The results first regards ROC curves and then turn to the actual control of the false alarm rate.

ROC comparison

Under \mathcal{H}_0 the noise is the AR(6) process described in (3.48). Under \mathcal{H}_1 , we consider the case of $N_S = 5$ sinusoidal signals with frequencies falling into a “valley” region of the noise PSD (cf. Fig. 3.12, right panel). The amplitudes of the sinusoids are all equals with $\alpha_q = 0.07 \text{ m}\cdot\text{s}^{-1}$, $\forall q = 1, \dots, 5$ and the frequencies are $f_q = [4, 4.5, 5, 5.5, 6] \text{ mHz}$.

This section compares the performance of the tests based on standardization of \mathbf{P} by a parametric estimate, $\hat{\mathbf{S}}_{E,AR}$, or by the non parametric estimate $\bar{\mathbf{P}}_L$ as in (3.2). The coefficients of the parametric noise PSD estimate, $\hat{\mathbf{S}}_{E,AR}$, are estimated by the best order and AR coefficients found on a time series of length $N \times L$ by the selection criteria described in (3.43). In the simulations, the best order is searched in the interval $[1, o_{max} = 40]$.

Fig. 3.16 shows the empirical ROC curves obtained through 10^4 MC simulation for five test statistics:

- Two are based on the standardized periodogram calibrated by a non parametric PSD estimate: $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ (black solid line) and $T_C(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ (red dashed solid line).
- Two are based on the standardized periodogram calibrated by the parametric PSD estimate as described above: $T_C(\tilde{\mathbf{P}}|\hat{\mathbf{S}}_{E,AR})$ with o estimated by FEP (blue points) and by MDL (yellow points).
- A last test statistics $T_C(\tilde{\mathbf{P}}|\hat{\mathbf{S}}_E)$ is added to illustrate the case where the $\hat{S}_{E,AR}$ coefficients are estimated with the true AR order $\hat{o} = o = 6$ here (green points). This case is an ideal (“oracle”) case for the standardization based on a parametric approach.

Each panel of Fig. 3.16 corresponds to one value of L (in reading order resp. for $L = 1, 20, 50$ and 100).

In the case $L = 1$, the “parametric tests” are more powerful than their non parametric counterparts, as AR estimation acts as a relevant prior knowledge. In practice, this means that the “whitening” effect of $\hat{S}_{E,AR}$ is efficient because the considered AR estimation is efficient. The ROC of the tests using the MDL criteria (which finds most of the time the true order in our example, cf. Fig. 3.4) and of the test using the effective true order are the best. Indeed, these tests have limited estimation error because the estimate $\hat{S}_{E,AR}$ is smooth, which is obviously not the case of the unconstrained, high variance estimate \bar{P}_1 (the variance of \bar{P}_L decreases in $1/L$ [Bartlett, 1950]). As L increases, the performance of parametric tests do not improve significantly. In contrast, those of the non parametric standardization increase with L (as the

variance at the denominator decreases) and the performance of all tests become comparable for $L \approx 20$. One can also see that, among the parametric tests, the test using the true model order has not the best performance and the test using the MDL criterion has the better ones.

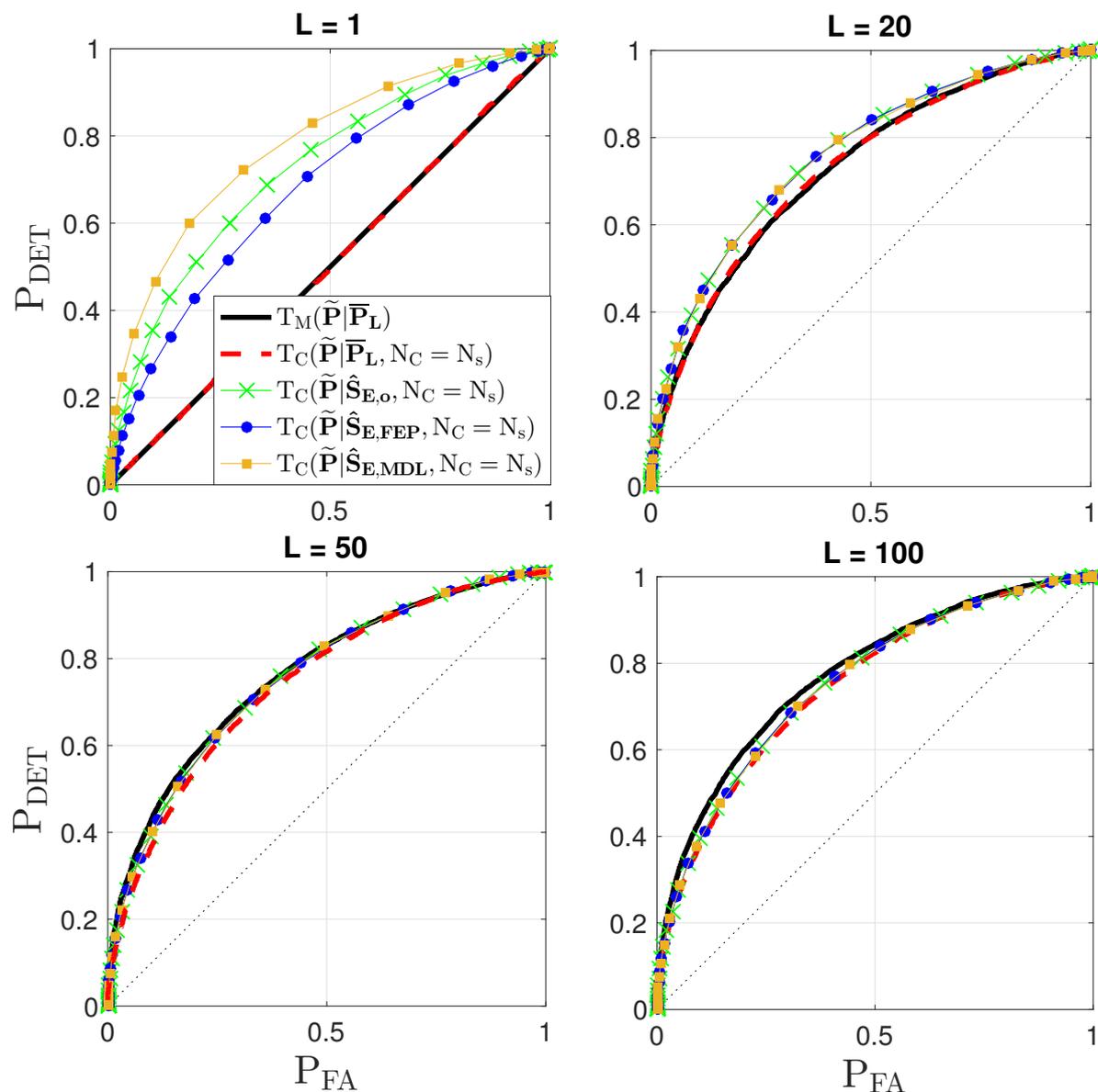


Figure 3.16 – ROC curves in case of signal frequencies falling into the PSD “valley” for different L values (1, 20, 50 and 100) associated to each panel (resp. from top left to bottom right). One can observe better performance for larger L for tests based on $\tilde{\mathbf{P}}|\overline{\mathbf{P}}_L$.

False alarm probability considering $\hat{S}_{E,AR} \approx S_E$

The parametric tests appear to have better performance in terms of ROC for small L . This is not the whole story, however. To see whether these tests can be implemented, we now investigate the relevance of the false alarm probabilities when derived under the approximation $S_E \approx \hat{S}_{E,AR}$ to compute target P_{FA} for tests based on $\tilde{\mathbf{P}}|\hat{S}_{E,AR}$. In the following study, we consider the P_{FA} of $T_C(\tilde{\mathbf{P}}|\hat{S}_{E,AR})$ approximated as in (3.47).

Fig. 3.17 compares, as a function of the test threshold, the P_{FA} (3.47) (blue curve) with the actual false alarm rates obtained for 1000 MC simulations. In each such simulation, an estimate $\hat{S}_{E,AR}(L)$ was obtained with the MDL criterion from L noise time series and used to calibrate the periodogram.

For each such estimate, the true P_{FA} of test $T_C(\tilde{\mathbf{P}}|\hat{S}_{E,AR}, N_C = N_s)$ was evaluated using 100 MC simulations and averages. The figure plots the averaged P_{FA} for $L = 1, 20$ and 100 (resp. the black, green and red solid lines with dots). One sees that (3.47) is accurate, *on average*, only when L is large. This figure indicates the variability of the true false alarm rate. For each value of L , the figure also shows the empirical 3σ dispersion of the true P_{FA} w.r.t. its empirical average (colored shaded regions). Even when N is large, the true significance levels at which T_C test is conducted undergoes wild (and in practice unknown) variations. The right panel is a zoom on the 3σ region for $L = 100$. Considering the threshold $\gamma = 5.9$ for instance, expression (3.47) leads to a supposed P_{FA} of ≈ 0.013 . In the right panel, one sees that the true false alarm rate for this threshold varies in reality in the range $[0.001, 0.03]$, depending on the realization of $\hat{S}_{E,AR}$. For smaller values of L , the excursions of the true false alarm rates are so large that (3.47) is simply useless (grey shaded regions). This shows that tests based on parametric estimation may be very hazardous, even when the parametric model is true and when large data sets are available for PSD estimation. Fig. 3.18 shows the empirical 1, 2, and 3σ envelopes for the three considered values of L .

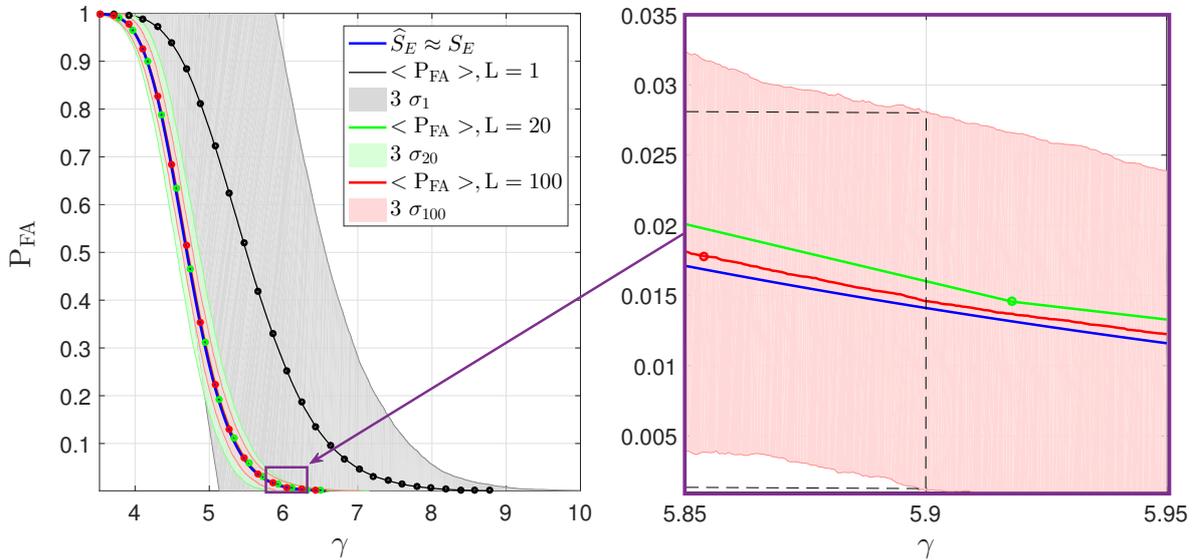


Figure 3.17 – Comparison of the approximated P_{FA} (3.47) (blue curve) with true P_{FA} for $L = 1, 20$ and 100. The solid lines represent the average actual P_{FA} . The shaded regions are the corresponding 3σ envelopes. The right panel is a zoom around $\gamma = 5.9$ (indicated by the purple boxed region in the left panel).

To conclude about the comparison between parametric *vs* non parametric approaches, the standardization by means of parametric estimates behave similarly for the different considered cri-

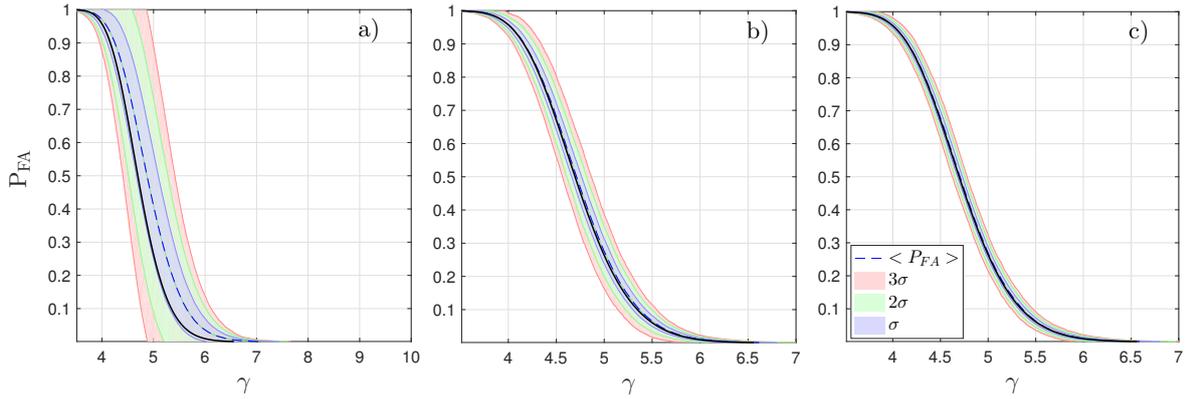


Figure 3.18 – Empirical 1, 2, and 3 σ envelopes of $P_{FA}(\gamma)$ curves for T_C applied to $\tilde{\mathbf{P}}|\hat{\mathbf{S}}_{E,AR}$ for different L : a) $L = 1$, b) $L = 20$, and c) $L = 100$.

teria, and they lead to better power than standardization using \bar{P}_L at low values of L . However, they do not allow to accurately evaluate the false alarm rate, even in the ideal configuration when the colored noise is generated according to the assumed parametric model. In contrast, standardization by \bar{P}_L allows to do so, with power similar to those of parametric methods when L reaches a few tens.

3.5.5 Application to the design of observational strategies

A direct application of the derived analytical expressions of the tests' probabilities of false alarm and detection is the design of observational strategies in the case of exoplanet detection in RV time series.

To illustrate this, we consider under \mathcal{H}_1 a planet with the parameters of α Centauri B's exoplanet candidate as estimated in [Dumusque et al., 2012] (see the legend of Fig. 3.19). As the orbital eccentricity of this planet is assumed null in this paper, the corresponding RV signal is very sparse in the Fourier domain. Hence, we consider here test T_M in (3.11) with P_{FA} given in (3.18) and P_{DET} in (3.19). The time sampling Δt is set to 4 hours but it was allowed to slightly vary from one value of N to another in order to guarantee that the planet's period yields a frequency exactly on the Fourier grid. In this case, the spectrum is 1-sparse in Ω and T_M is the test that yields the best performance.

Under \mathcal{H}_0 , one uses a model based on HD simulations of the Sun (cf. Perspective part). We note that there is some mismatch between the simulated RV time series of the solar spectral type (G2) and the true spectral type (K1) of the star hosting α CenB. Because spectral type affects the noise properties (e.g. see [Meunier et al., 2017a]), these results should not be considered to reflect perfectly the case of the candidate planet orbiting α CenB.

Fig. 3.19 illustrates the achievable performance compromises (P_{DET}, P_{FA}) as a function of N , for three target P_{FA} (0.5, 0.1 and 0.01, indicated respectively by the dotted, solid and dashed lines) and for different sizes of available training data sets L ($\infty, 100, 20, 5$, shown respectively in black, green, blue and red). These power curves were built using the expressions (3.21) and (3.19) for the T_M test and we checked that they are accurate in separate MC simulations (not shown). In the case $L \rightarrow +\infty$, we used the fact that $F(2, 2L) \xrightarrow[L \rightarrow +\infty]{d} \chi_2^2$ to calculate the theoretical $P_{DET}(P_{FA})$.

The study presented in Fig. 3.19 allows to quantify interesting facts. First, of course, P_{DET} is larger if the allowed P_{FA} is larger. Second, for a fixed P_{FA} , P_{DET} is larger for an increasing

values of L . Going to specific cases, one sees that if a planet similar to the considered candidate was orbiting a star of the considered spectral type (G2), of which 100 training time series are available, it would require 250 days ($1500 \times 4\text{h}$) of observations with 1 point every 4 hours to guarantee a probability of detection of 90% while ensuring a false alarm rate of 1%. This situation is indicated by the black square. With only $L = 5$ training time series, the probability of detection would fall to about 10%, all other parameters equals (red square).

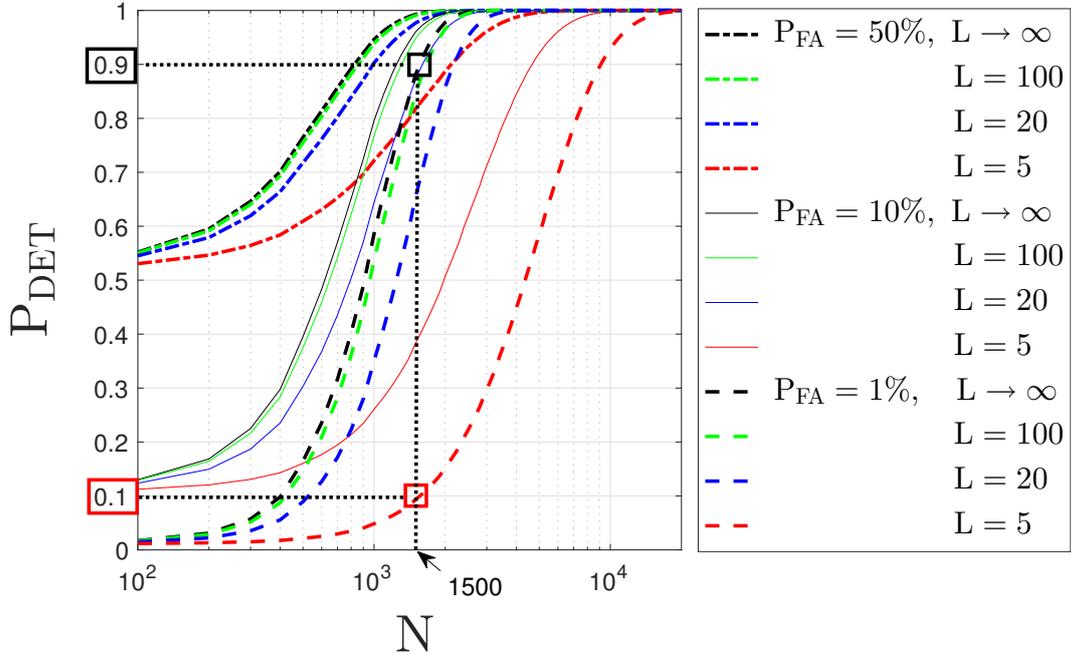


Figure 3.19 – Example of a detectability study for a planet orbiting a solar type star (relevant Keplerian parameters: $K_p = 0.54 \text{ m.s}^{-1}$, $T_p = 3.23 \text{ days}$, $e_p = 0$, $M_p = 1.241 M_\oplus$, $i = 90^\circ$, $d_p = 0.0425 \text{ AU}$). The plot shows the achievable P_{DET} for different configurations of P_{FA} budgets and numbers of available training light curves of the convective solar activity.

In this study, the frequency search is done on all the Fourier frequencies, i.e. in space Ω . However, one can note that we can obtain higher P_{DET} (at same P_{FA}) by considering a narrow frequency interval, i.e. $\Omega' \subset \Omega$.

These results show that, if a reliable set of L noise time series is available, our theoretical results on $T_M(\hat{\mathbf{P}}|\hat{\mathbf{P}}_L)$ (for instance) can be used to determinate the number of time needed to warrant any detection rate for a fixed P_{FA} .

3.6 Summary and conclusions

In this chapter, we have investigated the possibility of using training noise data sets to improve the reliability of the claimed false alarm rates. We have proposed to use these time series to standardize the periodogram and analyzed the performance of several tests applied to this periodogram.

In Sec. 3.2, we proposed an asymptotic analysis of the periodograms statistics after standardization in (3.2) for a model involving an unknown number of sinusoids with unknown parameters in partially unknown colored noise, i.e. model (3.1). This model is quite general and can be useful for a large variety of (quasi-) periodic signal detection problems.

In Sec. 3.3, this analysis allowed to characterize the performance of some standardized tests

in terms of false alarm and detection rates. The considered tests include classical approaches (T_M , T_C , T_F , T_{Ch} and $T_{F,rob}$) and also more recently considered adaptive tests (HC, BJ), designed for the rare and weak setting.

We showed that when standardization is performed with a simple averaged periodogram as a noise PSD estimate, tests as T_M , T_C , HC and BJ are CFAR for all sizes of training data sets. Moreover, FA and detection probabilities can be derived by analytical expressions for T_M and T_C . For adaptive tests, the standardization (by $\bar{\mathbf{P}}_L$) offers the same benefits as if the statistics of the noise were known *a priori*, with the CFAR property added. The false alarm and detection probabilities can not be derived analytically and the P -values involved in HC and BJ have to be computed through MC simulations or algorithms given in [Moscovich et al., 2016]. We also showed the statistical distributions of tests as T_F , T_{Ch} , and $T_{F,rob}$ applied to the standardized periodogram are not known when standardization with $\bar{\mathbf{P}}_L$ is involved (at least for finite values of L). In any case, without the standardization of \mathbf{P} as proposed in (3.2), the P_{FA} associated to any test statistics can not be derived when S_E is not known. In particular, we pointed out that standardization based on parametric estimates of the noise PSD exhibits actual false alarm rates that may be very far from the assumed ones, even for large data sets (cf. Sec. 3.5.4).

Regarding now the extensive numerical study of Sec. 3.6, one can list the following main results:

- Analytical P_{FA} and P_{DET} expressions derived in (3.18) and (3.19) for test $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ and derived in (3.38) and (3.41) for test $T_C(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ are reliable for finite values of N (when the correlation timescale is smaller than N). The influence of the noise correlation is then negligible, and the term ρ_k in (3.5) can be approximated to 1 in general. When the correlation timescale is large w.r.t. to the observation length N , the non-vanishing bias and dependencies at finite N are the main cause of mismatch between theoretical and empirical expressions (cf. Sec. 3.5.1).
- The performance of all test statistics based on $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$ logically increase with the number of available training data sets (cf. Sec. 3.5.1).
- T_C and the adaptive tests HC and BJ outperform T_M when the number of pure sinusoids under the alternative is large (while remaining small w.r.t. N). Adaptive tests can present better power than some procedures for which the number of sinusoids would be known in advance (cf. Sec. 3.5.2).
- Methods ignoring the noise correlations or using parametric noise PSD estimates can lead to highly unreliable estimates of the false alarm probability. For the latter case, this is even true when the noise model is exact, owing to the estimation noise of the model's parameters intrinsically injected in the detection procedure (cf. Sec. 3.5.3 and 3.5.4).
- Finally, the theoretical results on tests based on $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$, for which we can control the P_{FA} and P_{DET} , allow to design interesting detection strategies. These strategies can be computed for different signal parameters as well as observational constraints. For example, regarding the particular case of RV detection, such detection performance plots can be interesting when considering a regular time sampling of 1 point/night. These results allow to study best detection strategies in presence of convective noise, assuming of course that HD simulations of such noise are available to implement $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$.

In practical applications, some of the assumptions made in this chapter may not be met as for instance

- (i) the stationarity of the colored noise process,
- (ii) a regular sampling,
- (iii) a large N .

This is, in particular, the case in exoplanet detection using RV time series, owing for instance to astrophysical effects linked to magnetic activity (like spots, affecting (i)) or instrumental defects / observational constraints (affecting (i), (ii) and (iii)). Comparing theory to practice in this particular application of RV planets detection, the present study is nevertheless useful to make feasibility studies, which then describe best achievable performance (*i.e.*, around “quiet” stars and in absence of other un-modelled perturbations) for a regular sampling.

An important extension to the considered framework regards the case when, for periodicity analysis, the time series is not correlated to orthogonal exponentials. This case encompasses situations where (a) the sampling is irregular, (b) P is not evaluated on the Fourier grid (as in oversampled periodograms) (c) P is modified in the form of “generalized periodograms”, which correlate the time series with highly redundant dictionaries of specific features (e.g. [Bölviken, 1983b, Baluev, 2008, Süveges, 2014, Scargle, 1982, Bretthorst, 2003, Thong et al., 2004, Zechmeister and Kürster, 2009, Baluev, 2015, Gregory, 2016]). In such cases, the considered ordinates exhibit strong dependencies. With the additional complication of partially unknown colored noise, analytical evaluation of the false alarm rate for the considered tests seems out of reach. Addressing such situations is the purpose of the next chapter.

Chapter 4

Extension to the case of uneven sampling

Contents

4.1	Effect of uneven sampling on the periodogram	120
4.1.1	Considered sampling patterns	120
4.1.2	Classical periodogram and uneven sampling	123
4.2	Variants of the periodogram	123
4.2.1	The Lomb-Scargle periodogram	123
4.2.2	Least-square fitting and other related periodograms	126
4.3	Generalized Extreme Value distribution	128
4.3.1	Generalities	128
4.3.2	Parameters estimation	129
4.3.3	Diagnostic plots	130
4.4	False alarm rates evaluation	132
4.4.1	Method of “independent” frequencies	132
4.4.2	Bootstrap methods	133
4.4.3	FA evaluation using theory of the extremes	134
4.4.4	Conclusions	135
4.5	Proposed bootstrap procedure	136
4.5.1	Direct bootstrap approach	136
4.5.2	“Accelerated” bootstrap approach	141
4.6	Numerical studies on a synthetic noise process	141
4.6.1	Validation on known processes	141
4.6.2	“GEV accelerated” algorithm	148
4.7	Conclusions	151

This chapter investigates the difficult problem of assessing reliable false alarm rates for detection tests when, in addition to the presence of a colored noise with partially unknown statistics, the data sampling is irregular.

When the sampling is irregular, the classical periodogram ordinates $P(\nu_k)$ are not independent, even when the noise is uncorrelated (i.e. WGN), because the sampled exponentials in (2.24) are not orthogonal in general. This makes difficult to obtain analytical results for the P_{FA} (and the P_{DET}) of any test based on the periodogram ordinates. For instance, the analytical expressions of P_{FA} of tests T_M and T_C applied to $\mathbf{P}|\bar{\mathbf{P}}_L$ (3.2) (given resp. in (3.18) and (3.38)), are not valid for an uneven sampling.

In the aim of working with periodicity analysis tools, similar to the classical periodogram, different periodograms (mainly based on the Lomb-Scargle (LS) periodogram [Lomb, 1976, Scargle, 1982]) have been developed. Some of these periodograms variants are reviewed in Sec. 4.2.

In Sec. 4.3, we will introduce basic concepts of Generalized Extreme Value theory, which is the asymptotic distribution of extrema of r.v. This section will be used in the proposed bootstrap approach introduced in Sec. 4.5.

In Sec. 4.4, we will review some methods of the literature that aim to estimate the FA rates of tests based on LS periodogram using the assumption that some periodogram ordinates are independent or by using bootstrap procedures, or based on the theory of extreme values. As we will see however, all these methods have been developed for the case where the noise is White Gaussian.

Our contributions here are the following. Assuming again that we can dispose of reliable noise time series (e.g. through hydrodynamical simulations of the stellar activity), we propose in Sec. 4.5 an original and automated bootstrap method to produce reliable estimates of the P_{FA} of tests based on standardized periodograms of the form of (3.2).

The performance of such methodology to derive accurate false alarm rates will be evaluated numerically in Sec. 4.6 on a synthetic colored noise process, as done in the previous chapter. The investigation of the proposed methodology on real data and using real 3D hydrodynamical simulations of the stellar convective surface activity is one of the main perspective of this thesis.

The main results described in this chapter are published in [Sulis et al., 2017a].

4.1 Effect of uneven sampling on the periodogram

Let us first introduce the three main kinds of samplings that can be encountered in time series.

4.1.1 Considered sampling patterns

In the following, we consider different time samplings. They will be illustrated in Fig. 4.1 on a pure sine function of the form

$$R(t_j) = \alpha_s \sin(2\pi f_s t_j),$$

with $\alpha_s = 1$ and $f_s = 0.23 \mu\text{Hz}$. In this example, the initial (regular) time series length is $N = 100$ and $\Delta t = 1$ day. Each sampled signal is obtained by convolution of a continuous signal by a stream of Dirac impulsions. In each case, the spectral window¹ of the observation (cf. Sec. 2.2.8) is computed using zero padding with $N' = 10 \times N = 1000$ to better see the secondary peaks of the “continuous spectral window”. Moreover, the spectral windows will be normalized by their highest peak (in $\nu = 0$) to compare their different structures. In each case, the classical periodogram described in (2.24) is evaluated at dates $\{t_j\}$.

¹Computed as the FT of the time series with zeros when data are missing and ones otherwise. In the case of a regular sampling, we have seen in Sec. 2.2 that the spectral window is related to the Fejèr kernel [Li, 2014].

Regular sampling (\mathcal{S}_1). This sampling configuration, where the time sampling Δt is constant and the dates defined by $t_j = j\Delta t$, $\forall j = 1, \dots, N$, has been considered in Chap. 2 and 3. This sampling allows to define a “natural” frequency set for the periodogram defined by the Fourier frequencies given in (2.22). Associated to \mathcal{S}_1 , one can also define precisely the Nyquist frequency (f_{Ny}), i.e. the limit beyond which all signal frequencies (if present) will be aliased:

$$f_{Ny} := \frac{1}{2\Delta t}.$$

The first row of Fig. 4.1 illustrates a regular sampling of a pure sine $R(t_j)$ of frequency on Fourier grid ($f_s = \nu_{53}$, left), the spectral window of the observation (middle) and the periodogram (right) related to the convolution of the signal spectrum by the spectral window. In the spectral window, the zero padding allows to see the secondary peaks, which are null at the Fourier frequencies. In the periodogram, the convolution results in the presence of 2 peaks at frequencies $\nu_k = \pm f_s$ (cf. discussion in Sec. 2.2).

Regular sampling with gaps (\mathcal{S}_2). This configuration contains regularly sampled observations (e.g. in Astronomy, 1 point/night) where some data are occasionally missing (e.g. due to meteorological events) or periodically missing (e.g. due to seasonal events).

This case is illustrated in the second row of Fig. 4.1, where the pure sine function $R(t_j)$ (left, dots) is observed at three regular time intervals and contains two periodic gaps of 35 points each ($N_2 = \frac{30}{100}N = 30$). In the corresponding spectral window (middle), one can observe several secondary lobes at non null frequencies, which are multiples of the data gaps. In this case, the convolution of the signal spectrum by the spectral window produces additional lines in the periodogram as well (right) [Vityazev, 1994]. This can make the interpretation of the periodogram delicate (see e.g. the discussion in Chap.1 of [Bourguignon, 2006]).

Random irregular sampling (\mathcal{S}_3). In this configuration, the data samples are taken randomly in time. This case is illustrated in the last row of Fig. 4.1 where the sine function $R(t_j)$ (left, dots) is observed at irregular time intervals ($N_3 = N_2$ points selected randomly, here taken on the regular grid \mathcal{S}_1 , for a computational time reason). In the spectral window (middle), one can see several peaks, which appear as “noise” or “signal” peaks in the periodogram (right).

For signal detection, benefit could be taken from \mathcal{S}_3 , as this kind of irregularity allows to increase the number of frequencies free from aliases (as the alias’ origin comes from periodicities in the spectral window, see e.g. [Eyer and Bartholdi, 1999, Púčík et al., 2012]). A random time sampling improves on the aliases of the spectral window, thereby enhancing sparsity in the Fourier space under the alternative. We might thus expect that tests like T_M in (2.51), HC in (2.70) or BJ in (2.74) (applied to suitably normalized periodogram ordinates) would work better in irregular than in regular sampling. Regarding the results of our studies (cf. Sec. 4.5-4.6), the comparison of the performance of such tests in case of uneven and even samplings is straightforward and very interesting but is left in perspectives of this work.

For \mathcal{S}_3 , the definition of a Nyquist frequency is difficult. Bretthorst proposed to use a critical time, say δt , defined as the largest common divisor of all sampling epochs: $t_j := j\delta t$, with $j = 1, \dots, N' \in \mathbb{N}$ and $N' \gg N$ [Bretthorst, 1988]. If such a divisor exists, Eyer & Bartholdi showed that the spectral window periodicity is $1/\delta t$ and the Nyquist frequency can then be classically defined as $f_{Ny} = 1/2\delta t$ (e.g. this was the case for the numerical example in Fig. 4.1, cf. last row) [Eyer and Bartholdi, 1999]. If such a divisor does not exist, the Nyquist frequency does not exist as well. In practice, δt always exists for finite preci-

sion calculation \mathcal{S} but, if δt is too small, it may lead to computationally unfeasible calculations².

Regarding the application of interest, astronomical observations from the ground are affected by the day/night intermittence, the meteorological events or still instrumental failures during the data acquisition. Coupled with that, the observational time allocated to the researchers on high resolution instruments, necessary to detect the smallest planet signatures (such as HARPS, cf. Chap. 1), is difficult to obtain. Hence, gaps within the data of the form of \mathcal{S}_3 are common in astronomy and especially for ground based radial velocity surveys. This chapter is then an indispensable study if we want to apply our methodology on real RV observations.

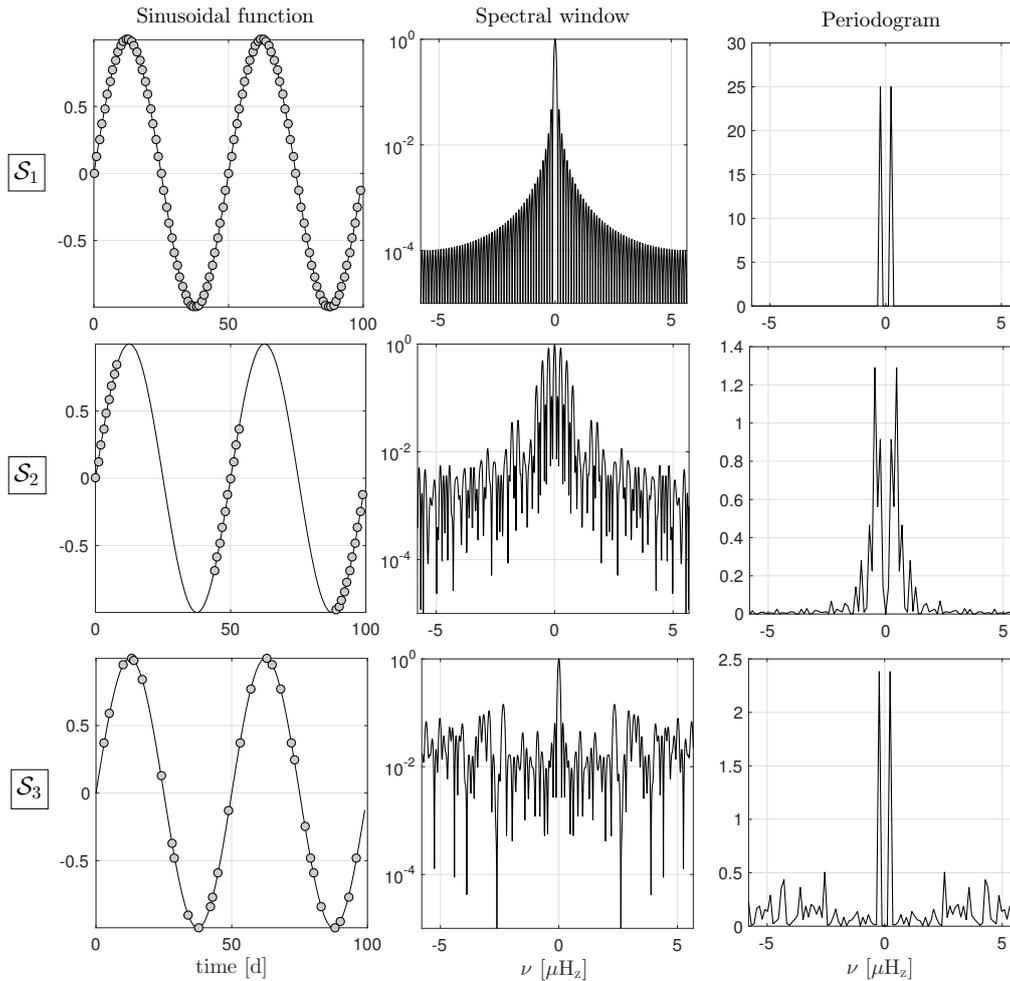


Figure 4.1 – Illustration of a sine function (column 1) in the Fourier domain (column 3) depending on the spectral window of the observations (column 2). The first row represents the case of a regular time sampling \mathcal{S}_1 ($N_1 = N = 100$, $\Delta t = 1$), the second row of \mathcal{S}_2 (regular sampling with two periodic data gaps of 35 missing points each, $N_2 = 30$), and the last row illustrates a random irregular sampling \mathcal{S}_3 ($N_3 = 30$). When gaps are present, one can observe peaks at frequencies $\nu \neq f_s$ in the periodograms due to the spectral window of the observations (cf. two latter rows). These peaks look like noise or signal peaks in the data.

² Note that different methods have been proposed in the literature to estimate the Nyquist frequency: some consider the mean of the sampling interval [Scargle, 1982, Horne and Baliunas, 1986], the average of the inverse time intervals between the measurements [Debosscher et al., 2007], the median [Graham et al., 2013] or the minimum sampled spacing [Roberts et al., 1987]. However, while all the limit frequencies defined by these different techniques asymptotically tend to the definition of f_{Ny} for \mathcal{S}_1 , none of these approaches seems strictly valid in practice (e.g. see the discussion p.18 of [VanderPlas, 2017]).

4.1.2 Classical periodogram and uneven sampling

When the sampling is regular (\mathcal{S}_1), the classical periodogram ordinates in (2.24) can be interpreted as the energy obtained when projecting the data on orthogonal exponentials. Indeed, as shown in Chap. 2, P can be rewritten in this case as:

$$\begin{aligned} P(\nu) &:= \frac{1}{N} \left| \sum_{j=0}^N X(t_j) e^{-2\pi i t_j \nu} \right|^2 \\ &\stackrel{\mathcal{S}_1}{=} \left| \langle \mathbf{X}, \mathbf{g}_j \rangle \right|^2 \quad \text{with } \mathbf{g}_j := [e^{-2\pi i j \nu}, \dots, e^{-2\pi i j N \nu}]^\top \\ &\stackrel{\mathcal{S}_1}{=} \frac{1}{N} \left[\left(\sum_{j=0}^N X(t_j) \cos(2\pi i t_j \nu) \right)^2 + \left(\sum_{j=0}^N X(t_j) \sin(2\pi i t_j \nu) \right)^2 \right]. \end{aligned} \quad (4.1)$$

When the sampling is irregular ($\mathcal{S}_2, \mathcal{S}_3$), the complex exponentials \mathbf{g}_j are not orthogonal (nor have the same norm), so $\langle \mathbf{g}_j, \mathbf{g}_k \rangle \neq 0$ for $j \neq k$, and the set of periodogram ordinates are correlated (and thus dependent). It is also possible to form “generalized periodograms”, which correlate the time series with possibly highly redundant dictionaries of specific features. The most famous (and used) such periodogram is the *Lomb-Scargle periodogram* [Lomb, 1976, Scargle, 1982].

4.2 Variants of the periodogram

4.2.1 The Lomb-Scargle periodogram

Lomb & Scargle have shown that the correlation between the sine and cosine terms in (4.1) can be cancelled by construction of a new periodogram [Lomb, 1976, Scargle, 1982]. For WGN, the two terms in (4.1) are two χ_1^2 and hence this periodogram still has a χ^2 distribution.

More specifically, Scargle addressed a general form for the classical periodogram written as

$$P(\nu) = \frac{A^2}{2} \left(\sum_{j=0}^N X(t_j) \cos(2\pi\nu[t_j - \tau(\nu)]) \right)^2 + \frac{B^2}{2} \left(\sum_{j=0}^N X(t_j) \sin(2\pi\nu[t_j - \tau(\nu)]) \right)^2 \quad (4.2)$$

with $A, B \in \mathbb{R}$ and $\tau(\nu)$ an arbitrary function of the frequency and time epoch to be optimized. The value of the phase is chosen such the periodogram in (4.2) (i) reduced to its classical form for regularly sampled observations (leading to a χ^2 distribution), and (ii) is invariant to a time-shift of the time epochs (as the classical periodogram in (2.24) for a regular sampling). The resulting Lomb-Scargle (LS) periodogram is easy to implement and defined as:

$$P_{LS}(\nu; A, B, \tau) := \frac{1}{2} \left(\frac{\left(\sum_{j=1}^N X(t_j) \cos(2\pi\nu[t_j - \tau]) \right)^2}{\sum_{j=1}^N \cos^2(2\pi\nu[t_j - \tau])} + \frac{\left(\sum_{j=1}^N X(t_j) \sin(2\pi\nu[t_j - \tau]) \right)^2}{\sum_{j=1}^N \sin^2(2\pi\nu[t_j - \tau])} \right) \quad (4.3)$$

with

$$\tau(\nu) := \frac{1}{4\pi\nu} \arctan \left(\frac{\sum_{j=1}^N \sin(4\pi\nu t_j)}{\sum_{j=1}^N \cos(4\pi\nu t_j)} \right). \quad (4.4)$$

The frequency ν in (4.3) and (4.4) is usually taken in the interval $\nu \in [0, (N - 1)\Delta\nu]$ with $\Delta\nu = \frac{1}{T}$ (but, as discussed in Sec.4.1.1, the frequency interval can be chosen differently). The LS periodogram differs from the classical one by the denominators in (4.3) (namely, $\sum_j \cos^2(2\pi\nu[t_j - \tau])$ and $\sum_j \sin^2(2\pi\nu[t_j - \tau])$) to be compared to $N/2$ in (2.24) (cf. Fig.5 of [Scargle, 1982]). We note, however, that in many cases of irregular sampling, P_{LS} differs only slightly from the classical periodogram (cf. Example. 14). Moreover, the ordinates of P_{LS} are not independent for uneven samplings \mathcal{S}_2 and \mathcal{S}_3 , even for WGN.

Example 14. Comparison between classical and LS periodograms for \mathcal{S}_1 and \mathcal{S}_2

Consider a model involving under \mathcal{H}_0 a WGN $\overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ of (known) unit variance and under \mathcal{H}_1 a sine function of parameters $\alpha_s = 1 \text{ m.s}^{-1}$ and $f_s = 0.2 \text{ Hze.g.}$. The time series are irregularly sampled on $N' = 50$ data points (N' points are randomly taken in a larger regular grid of $N = 500$ points).

Panel a) of Fig.4.2 shows one realization of the classical (black) and LS (red) periodograms under \mathcal{H}_1 . The classical periodogram has been computed on a regular grid with $\Delta t = 1 \text{ s}$ and missing data are treated as zeros. Panel c) shows the absolute difference as a function of the frequency between these two periodograms. Panel b) shows the empirical ROC curves of test T_M given in (2.51) applied to both periodograms. These curves have been computed using 10^5 MC simulations of the considered model. Panel d) represents the difference in terms of P_{FA} and P_{DET} by computing their absolute difference at each test threshold γ .

One can see that, in terms of ROC, test T_M based on the LS periodogram has better (though not significantly better) detection performance (see [Vio et al., 2010] for similar conclusions).

In case of uncorrelated noise $X(t_j) \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_j^2)$ under the null (with possibly σ_j^2 different but known), Scargle has shown that the LS periodogram in (4.3) is equivalent, at each considered frequency ν , to a data fitting by a sinusoidal function of the form

$$\mathbf{Y}_{LS}(\mathbf{t}_j; \boldsymbol{\theta}_{LS}) = A \cos(2\pi\nu t_j + \tau(\nu)) + B \sin(2\pi\nu t_j + \tau(\nu)), \quad (4.5)$$

with $\boldsymbol{\theta}_{LS}(\nu) = [A, B, \tau(\nu)]^\top$ the unknown model's parameters. The P_{LS} can then be seen as a least-square fitting procedure (“ χ^2 minimization procedure”), i.e. a function that minimises the mean square difference between the data and a model function of the form:

$$f_{LS}(\mathbf{t}; \boldsymbol{\theta}_{LS}) := \sum_j \frac{(X(t_j) - Y_{LS}(t_j; \boldsymbol{\theta}_{LS}))^2}{\sigma_j^2}, \quad (4.6)$$

From this point of view, the LS periodogram is very similar to the FT in the classical periodogram: sine and cosine function are fitted across a grid of frequencies.

When the noise is white Gaussian, one can show that each individual P_{LS} ordinate is χ^2 distributed, $\forall \nu$ [Scargle, 1982]:

$$P_{LS}(\nu) \sim \varphi_{\chi^2_2}(\gamma) = \frac{1}{\sigma^2} e^{-\gamma/\sigma^2}. \quad (4.7)$$

The cdf of (4.7) is of the form

$$\Phi_{P_{LS}}(\gamma) = 1 - e^{-\gamma/\sigma^2}. \quad (4.8)$$

Consequently, σ^2 in (4.7) and (4.8) has to be known if we want to be able to set the threshold γ for a target P_{FA} in Schuster's test for instance.

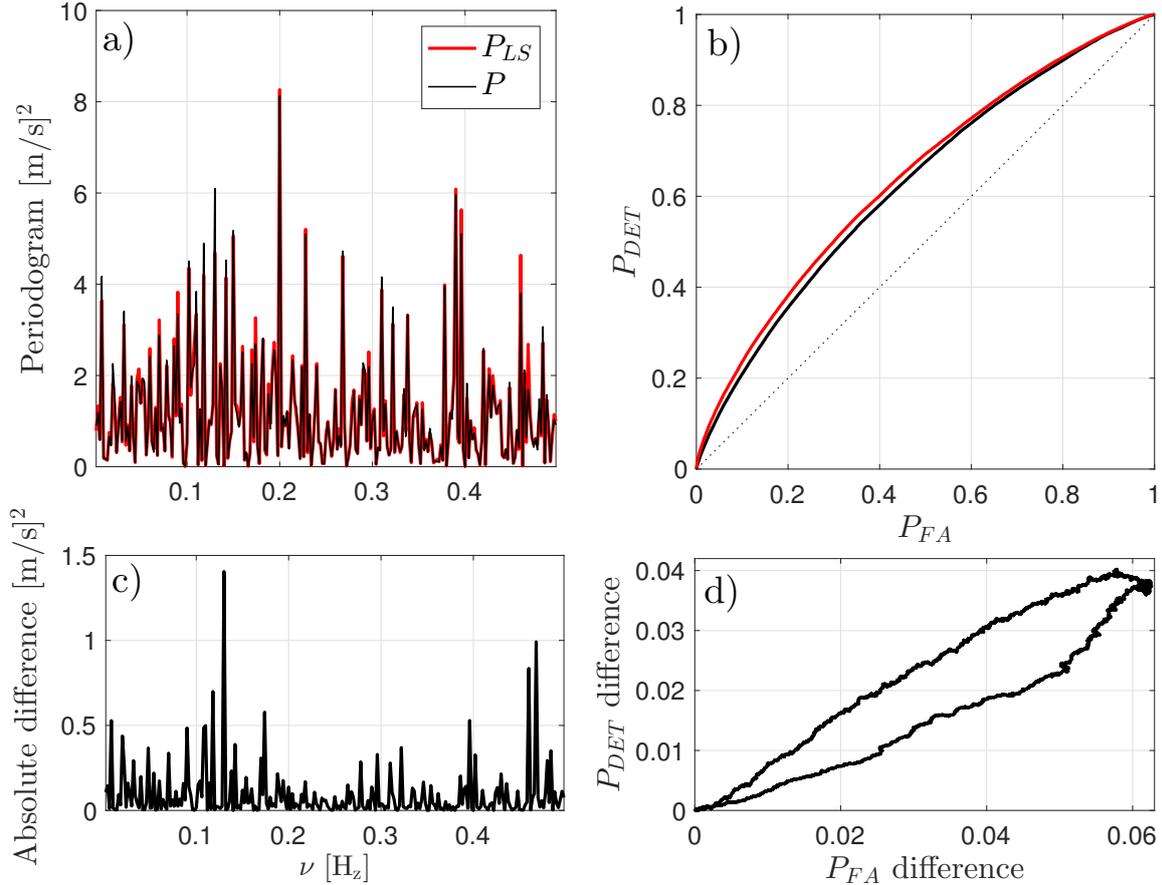


Figure 4.2 – Comparison between classical (black) and LS (red) periodograms, given in (2.24) and (4.3), for series of sine in WGN unevenly sampled. The signal frequency is 0.2 Hz. a) One periodogram realization, b) ROC comparisons, c) absolute periodogram difference at each frequency, and d) absolute ROC difference computed at each threshold γ . One can observe that the periodograms (and associated ROC) differ only slightly.

When the noise variance is unknown, different periodogram normalizations exist (e.g. [Schwarzenberg-Czerny, 1998, Baluev, 2008]), but they seem to be rarely used in practice.

When X is an unknown colored noise, the statistical distribution of (4.3) is unknown and the interpretation of P_{LS} has to be done carefully [Vio et al., 2010]. In general, classical procedures consist in

1. Computing the P_{LS} ,
2. Whitening all the peaks assumed caused by the noise through a procedure such as CLEAN³ [Roberts et al., 1987],
3. Computing P_{LS} on the data residuals,
4. Standardizing the P_{LS} by the estimated variance of the data residuals, say $\hat{\sigma}_{res}^2$: $\frac{P_{LS}}{\hat{\sigma}_{res}^2}$,
5. Estimate the P_{LS} residual peaks significance assuming the data residuals are uncorrelated, i.e. $\frac{P_{LS}}{\hat{\sigma}_{res}^2} \sim \chi_2^2$.

³In a few words, a CLEAN procedure consists in removing iteratively the largest peak in the periodogram through a fit of a given function to the data (in general, a sinusoid with frequency given by the considered peak location).

The last point is only approximately correct at each frequency and the marginal joint distribution of the periodogram ordinates and their intrinsic dependences remain unknown.

Before discussing the control of the FA rate in case of colored noise for uneven sampling, which is of concern in this thesis, let us review other interesting variants of the classical periodogram (2.24).

4.2.2 Least-square fitting and other related periodograms

In the literature, one finds different forms of model \mathbf{Y}_{LS} in (4.5) involved in the minimization of a function of the form:

$$f_Y(\mathbf{t}; \boldsymbol{\theta}) := (\mathbf{X} - \mathbf{Y}(\mathbf{t}; \boldsymbol{\theta}))^\top \Sigma^{-1} (\mathbf{X} - \mathbf{Y}(\mathbf{t}; \boldsymbol{\theta})) \quad (4.9)$$

with $\mathbf{X} = [X_1, \dots, X_N]^\top$ the observed data of covariance matrix Σ and $\mathbf{Y}(\mathbf{t}; \boldsymbol{\theta}) = [\mathbf{Y}(t_1; \boldsymbol{\theta}), \dots, \mathbf{Y}(t_N; \boldsymbol{\theta})]^\top$ the signal model parametrized by $\boldsymbol{\theta}$.

For example, one can note the *floating mean periodogram* consisting in adding an offset constant $C \in \mathbb{R}$ to model Y_{LS} in (4.5) to makes (4.9) equivalent to a constant plus sine wave fitting [Cumming et al., 1999]:

$$\mathbf{Y}_C(t_j) = A \cos(2\pi\nu t_j) + B \sin(2\pi\nu t_j) + C. \quad (4.10)$$

The data-compensated DFT (DCDFT) considers a model \mathbf{Y}_{DC} of the form [Ferraz-Mello, 1981]

$$\mathbf{Y}_{DC}(t_j) = A \sin(2\pi\nu t_j) + B\epsilon(t_j), \quad (4.11)$$

with A an arbitrary constant, $\epsilon \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ and B a factor related to the SNR. The stochastic term of (4.11) has been introduced to counteract the fact that the mean of the data is not well estimated when the sampling is irregular. For example, a local concentration of data will overestimate the estimated mean in contrast to largely spaced irregular data. As for the P_{LS} , when the noise is a WGN, each ordinate of the DCDFT is χ^2 distributed but they are not independent [Bourguignon, 2006].

Zechmeister & Kürster defined a *Generalized LS periodogram* by introducing weight functions to the model \mathbf{Y}_C in (4.10) to take into account that some data are more reliable than others [Zechmeister and Kürster, 2009]. However, in contrast to the general case of (4.9), this measure of the data uncertainties takes into account the error at each time t_j (e.g. due to observational conditions as the seeing or temperature variations) but not the noise correlations. Consequently, considering a WGN, Zechmeister & Kürster write (4.9) as

$$f_Z(\mathbf{t}_j; \boldsymbol{\theta}) := W \sum_{j=1}^N w_j \left(X(t_j) - Y(t_j; \boldsymbol{\theta}) \right)^2 \quad (4.12)$$

with $w_j = \frac{1}{W} \frac{1}{\sigma_j^2}$ the normalized weights, $W = \sum_{i=1}^N \frac{1}{\sigma_i^2}$, and $\sum_{j=1}^N w_j = 1$.

Regarding the literature in astronomy for example, other periodogram variants exists:

Bayesian periodograms. *Bayesian Keplerian periodograms* involve a model, say Y_K , based on (1.3) [Cumming, 2004, Jenkins et al., 2014, Tuomi et al., 2014, Mortier et al., 2015]. They are an extension of the LS periodogram but with a Keplerian RV model, “weight” functions and a constant offset added to the data. The parameters space can be explored using different methods, which are generally computationally heavy (e.g. using adaptive Metropolis algorithm [Tuomi et al., 2014]).

Methods using Bayesian Keplerian periodograms were shown to be the most successful methods (compared to the other) to detect (by iteration) the synthetic planetary signals injected in noisy data while ensuring a low false positive detection rate in the International Challenge proposed by Dumusque [Dumusque et al., 2017]. In practice, Keplerian periodograms have been also successful to detect several planets (see e.g. the detection of a cold Saturn planet in [Tuomi et al., 2014], or the claimed 6 planets systems of GJ876 in [Jenkins et al., 2014]).

Maximum-Likelihood periodograms. Baluev proposed a *maximum-likelihood periodogram* (MLP) that involve the multi-planet Keplerian parameters fitting and uncorrelated Gaussian noise RV errors (method available through the software *PlanetPack*⁴, which can be used for the goal of exoplanets detection [Baluev, 2013b]). The MLP represents a likelihood function (2.14), which compares a given model to the data at each step of the minimization in (4.9). In the case of a Gaussian noise, the likelihood function writes:

$$\mathcal{L}(\mathbf{x}; \boldsymbol{\theta}_R, \boldsymbol{\theta}_E) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\boldsymbol{\theta}_E(i)}} \exp\left\{-\frac{[\mathbf{X}(t_i) - \mathbf{Y}_K(t_i; \boldsymbol{\theta}_R)]^2}{2 \xi \boldsymbol{\theta}_E(t_i)}\right\} \quad (4.13)$$

with \mathbf{Y}_K the considered Keplerian model, $\boldsymbol{\theta}_R$ the model parameters depending on the Keplerian parameters related to (1.3), $\boldsymbol{\theta}_E(t_i) := \sigma_{\star}^2(t_i) + \sigma_G^2$ the noise parameters depending on the stellar (σ_{\star}^2) and instrumental (σ_G^2) noise components to be estimated, and ξ a corrective factor depending on the used number of degrees of freedom (here, of the size of $\boldsymbol{\theta}_R$). Function (4.13) has to be maximized and the significance, according to the model, is directly evaluated from the obtained values. One of the main advantages of such a method is that it is able to detect multiple signals at the same time (i.e. it is not influenced by data removal since it is a joint estimation) and is opened to include correlated noise along with the WGN component.

Global fitting approaches. In the case of multiple signals (or a signal with harmonics), the classical procedure based on the LS periodogram consists in subtracting out the highest components before a new application of the method on the data residuals. This can induce problems however, as the removal, one by one, of the highest components affects the periodogram at all frequencies at each iteration. Hence, such iterative methodology is critical for detecting low mass planets as SNR are typically low and each iterative step can possibly influence the peaks due to the signal of interest [Jenkins et al., 2014]. To counteract this limit, “multi-sines” fitting methods, which consists in estimating simultaneously all the periodogram components, have been developed (e.g. [Bourguignon et al., 2007, Hara et al., 2017]).

For example, Hara et al. proposed to fit the multi planet signatures simultaneously during the detection procedure. Their method (called the *basis pursuit ℓ_1 -periodogram*) is a sparsity driven technique, where the observed RV are fitted using a global model fitting approach combined with a Gaussian process to account for the noise correlations. They applied their method to published detected multi-planetary systems and shared the first place, with the Bayesian Keplerian periodograms, in the International Challenge of Dumusque [Dumusque et al., 2017].

Alternatively, Feng et al., proposed an alternative Bayesian method, to separate in the same time both signals and correlated noises [Feng et al., 2017]. This avoids the bias introduced by any residual based analysis method. Their routine consists in fitting the correlated noise using a Moving Average (MA) model, comparing the noise models scores, selecting one, and optimizing a frequency-dependent linear trend simultaneously with the sinusoids and noise components. The significance of signals

⁴<https://sourceforge.net/projects/planetpack/>

are then accessed visually, through a *moving periodogram*, built using the Bayes' factor (estimated by the Bayesian information criterion, BIC, as described in [Feng et al., 2016]).

To summarize, there exists numerous interesting periodogram-like tools to deal with irregular sampling. The majority of these models have reached a lot of success in extrasolar planet detection.

However, these methods have also their drawbacks when dealing with signal buried in the stellar noise as this colored noise is difficult to be isolated from the planetary signal. The choice of the noise model can greatly influence the detection choice. For example, around the star GJ 581, a system of 6 planets has been found by using a WGN model [Vogt et al., 2010], while planets f and g were refuted when using a Gaussian red noise model (cf. [Baluev, 2013a], who also questioned the existence of planet d). As noticed in [Feng et al., 2016], this is also the case for the system around GJ 667C, which was claimed using WGN models and refuted by methods using MA and Gaussian models [Anglada-Escudé et al., 2013]. Moreover, complicated models (as the Keplerian periodograms) are computationally heavy (as a lot of parameters have to be considered), which can make their use complicated for long time series.

4.3 Generalized Extreme Value distribution

Before discussing the false alarm probability applied to periodograms such as the P_{LS} , we dedicate a section to an introduction to *Generalized Extreme Value* (GEV) theory, which will be useful for Sec. 4.4.3 and in our study in Sec. 4.5.2. This introduction is mainly based on the Coles' book [Coles, 2001].

4.3.1 Generalities

The GEV approach consists in looking at the asymptotic ($N \rightarrow \infty$) behavior of the distribution of the extremum of N r.v. (minima as well as maxima).

Consider here test T_M (2.51) applied to a vector \mathbf{Z} and the linear renormalization of $T_M(\mathbf{Z})$:

$$T_M^*(\mathbf{Z}) := \frac{T_M(\mathbf{Z}) - b_N}{a_N}, \quad (4.14)$$

with $a_N > 0$ and b_N some constants. The following theorem (extracted from [Coles, 2001], Theorem 3.1, p.46) gives the three possible (asymptotic) distributions of (4.14).

Theorem 3. (Extremal types theorem - Theorem 3.1. [Coles, 2001]).

If there exist sequences of constants $a_N > 0$ and b_N such that

$$\Pr\left(\frac{T_M(z) - b_N}{a_N} \leq z\right) \xrightarrow{N \rightarrow \infty} G(z),$$

where G is a non-degenerate function, then G belong to one of the following distributions:

$$\left\{ \begin{array}{l} I : G(z) = \exp\left\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right\}, \quad -\infty < z < \infty \\ II : G(z) = \begin{cases} 0, & z \leq b, \\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\}, & z > b, \end{cases} \\ III : G(z) = \begin{cases} \exp\left\{-\left[-\left(\frac{z-b}{a}\right)^\alpha\right]\right\}, & z < b \\ 1, & z \geq b. \end{cases} \end{array} \right.$$

for positive values of parameters a, b , and α .

This theorem is valid in the asymptotic regime, for all random variables z , which are sets of maxima or minima. The three distributions $G(z)$ of Theorem. 3 are (cf. [Katz, 2008] for the different examples):

- I- Gumbel distribution: the density of G decays exponentially, the upper end-point u is infinite. The probability that z achieve large maxima is lower than for Fréchet distribution, but it can happens anyway (contrary to the Weibull distribution). Examples are: normal (i.e. Gaussian), log-normal, Gamma, exponential and χ^2 distributions.
- II- Fréchet distribution: the density of G decays polynomially, the upper end-point u is also infinite. There is a non negligible chance than z achieve the extreme values. Examples are: Student-t and Cauchy distributions.
- III- Weibull distribution: the upper end-point u is finite. In this case, $z < \mu - \xi\sigma$ and there is zero probability to have value above this limit. Examples are: β and uniform distributions.

These three families of distributions will be illustrated in Example. 16. The distribution I and II correspond to different rate of decay in the tail of the distribution of $T_M(\mathbf{Z})$, which are parametrized by α .

The choice between these three distributions, as *a priori* distributions, is delicate. A better way to avoid this choice is to reformulate the extremal types theorem with a single family superseding the three distributions: the GEV distribution [Fisher and Tippett, 1928, Jenkinson, 1955]. The GEV distribution has cdf:

$$G(z; \boldsymbol{\theta}_{GEV}) := \begin{cases} \exp\left\{-\left[1 + \xi\left(\frac{z - \mu}{\sigma}\right)\right]_+^{-1/\xi}\right\}, & \text{for } \xi \neq 0 \\ \exp\left\{-\exp\left\{-\left(\frac{z - \mu}{\sigma}\right)\right\}\right\}, & \text{for } \xi = 0, \end{cases} \quad (4.15)$$

where the notation a_+ means $a_+ := \max(0, a)$ and $\boldsymbol{\theta}_{GEV} := [\xi, \mu, \sigma]^\top$ the GEV parameters vector with $\xi \in \mathbb{R}$ the shape, $\mu \in \mathbb{R}$ the location, and $\sigma \in \mathbb{R}^+$ the scale. In the case $\xi \neq 0$, (4.15) is defined for $z : 1 + \xi\frac{z - \mu}{\sigma} > 0$. In the case $\xi = 0$, (4.15) is derived by taking the limit in 0 and gives the Gumbel distribution. A value $\xi > 0$ corresponds to the Fréchet distribution, $\xi = 0$ to the Gumbel distribution, and $\xi < 0$ to the Weibull distribution.

An important quantity is the *return level function*, $\gamma(q)$, evaluated for a return period q . In common terminology, the return level is the level that is expected to be exceeded once in every q random trials under the same characteristics. For example the return level $\gamma(1/10)$ is expected to be exceeded on average once every 10 time series of same length and time sampling grid. The FA threshold associated to a target P_{FA} is simply the $1 - P_{FA}$ quantile of the GEV distribution ($G(\gamma; \boldsymbol{\theta}_{GEV}) = 1 - P_{FA}$), which can be found by inverting (4.15) for test $T_M(\mathbf{Z})$ in (4.14):

$$\gamma(P_{FA}; \boldsymbol{\theta}_{GEV}) = G^{-1}(1 - P_{FA}; \boldsymbol{\theta}_{GEV}) = \begin{cases} \mu - \frac{\sigma}{\xi} \left(1 - \left(-\log(1 - P_{FA})\right)^{-\xi}\right), & \text{for } \xi \neq 0, \\ \mu - \sigma \log\left(-\log(1 - P_{FA})\right), & \text{for } \xi = 0. \end{cases} \quad (4.16)$$

The demonstration of (4.16) is reported in Appendix. D.

4.3.2 Parameters estimation

The standard method to estimate the three unknown parameters $\boldsymbol{\theta}_{GEV}$ of the GEV distribution is to maximize the log-likelihood function associated to (4.15).

This log-likelihood function is easily obtainable from the product of the GEV density. Denoting the vector of b maxima $\mathbf{Z} = [Z_1, \dots, Z_b]^\top$ following the GEV distribution (say g), $Z_i \stackrel{i.i.d}{\sim} g(\boldsymbol{\theta}_{GEV})$, the log-likelihood writes:

$$\ell(\mathbf{z}; \boldsymbol{\theta}_{GEV}) = \begin{cases} -b \log(\sigma) - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^b \log\left(1 + \xi \frac{z_i - \mu}{\sigma}\right) - \sum_{i=1}^b \left(1 + \xi \frac{z_i - \mu}{\sigma}\right)^{-1/\xi}, & \text{for } \xi \neq 0, \\ -b \log(\sigma) - \sum_{i=1}^b \left(\frac{z_i - \mu}{\sigma}\right) - \sum_{i=1}^b \exp\left\{-\left(\frac{z_i - \mu}{\sigma}\right)\right\}, & \text{for } \xi = 0. \end{cases} \quad (4.17)$$

with $1 + \xi \frac{z_i - \mu}{\sigma} > 0$ for the case $\xi \neq 0$ (as otherwise (4.17) goes to $-\infty$) and b the number of considered maxima. The demonstration of (4.17) is reported in Appendix. D.

Maximizing this function consists in differentiating (4.17) with respect to each parameter $\boldsymbol{\theta}_{GEV}$. It yields a system of equations given in (D.8) (cf. Appendix. D). As it does not exist any analytical solution to solve these equations, they have to be solved iteratively using standard numerical algorithms such as the Newton-Raphson method. It gives the MLE parameters $\hat{\boldsymbol{\theta}}_{GEV} := [\hat{\xi}, \hat{\mu}, \hat{\sigma}]^\top$. The method to evaluate the standard errors on these GEV parameters is also reported in Appendix. D (cf. (D.17)).

4.3.3 Diagnostic plots

To verify the “quality” of the parametric GEV modelization, one can usually uses *diagnostic plots* as the *Quantile-Quantile* (Q-Q) or the *Return level* plots.

Quantile-Quantile plot The Q-Q plot aims to verify, visually, the validity of the GEV model by comparing the empirical distribution of the independent maxima with the GEV model. Considering the ordered maxima $z_{(i)}$, defined for $i = 1, \dots, b$, the empirical probability to have exactly i values less than or equal to $z_{(i)}$ is i/b . Slightly changing⁵ this probability by $\frac{i}{b+1}$, the Q-Q plot at the point $z_{(i)}$ writes (abscissas, ordinates):

$$\left\{ \hat{G}^{-1}\left(\frac{i}{b+1}; \hat{\boldsymbol{\theta}}_{GEV}\right), z_{(i)} \right\}, \quad \text{for } i = 1, \dots, b.$$

If the fitted GEV model is good

$$\hat{G}(z_{(i)}; \hat{\boldsymbol{\theta}}_{GEV}) \approx \frac{i}{b+1},$$

and the plot is close to the unit diagonal. Deviations from linearity are indicative of an invalid GEV model. Contrary to the histogram plots of the data samples, which are generally less informative as their shape is sensitive to the chosen samples grouping intervals (i.e. the bins), the Q-Q plot allows to see the possible divergences in the distribution tails (where we usually want reliable return level precision, cf. e.g. to the discussion in Sec. 2 of [Süveges, 2014]).

Return level plot The return level plot is the graph used to evaluate if the variations in the Q-Q plot are “natural” or in contrary if they are caused by an invalid GEV model. This plot consists in plotting the return level (given in (4.16)) against the probability values $[-\log(1 - q)]$ in logarithm scale:

$$\left\{ \log\left[-\log(1 - q)\right], \gamma(P_{FA}; \boldsymbol{\theta}_{GEV}) \right\} \quad \forall 0 < q < 1.$$

⁵To avoid the cdf value 1, which could cause problems as noted in [Coles, 2001, Süveges, 2014].

and to compare it with the maximum likelihood return level estimate $\hat{\gamma}(P_{FA}; \hat{\boldsymbol{\theta}}_{GEV})$ evaluated by replacing $\boldsymbol{\theta}_{GEV}$ by $\hat{\boldsymbol{\theta}}_{GEV}$ in (4.16). The abscissa are plotted in logarithm scale because it compresses the tail of the distribution and then displays the return level estimates for long return periods.

If $\xi = 0$ the plot is linear. If $\xi < 0$, the plot is convex with asymptotic limit as $p \rightarrow 0$ at $\frac{\sigma}{\xi}$, and if $\xi > 0$ the plot is concave and has no finite bound. Confidence intervals can be associated to the return level plots to be more informative (cf. (D.20) of Appendix. D).

Example 15. Illustration of the three GEV distribution families

To illustrate the three different families of the GEV distribution, we generate $b = 10^4$ realizations of i.i.d. r.v. X with length $N = 1000$. For each series, we compute the maximum, resulting in a vector $\mathbf{Z} = [Z_1, \dots, Z_b]^\top$ of maxima and the parameters $\hat{\boldsymbol{\theta}}_{GEV} := [\hat{\xi}, \hat{\mu}, \hat{\sigma}]^\top$ are evaluated as described in Sec. 4.3.2. The 95% confidence intervals are obtained using (D.17) (cf. Appendix. D). Three cases are studied:

- Case 1: the r.v. follow a Student-t distribution $X_i \stackrel{i.i.d.}{\sim} t(5)$ with 5 degrees of freedom,
- Case 2: the r.v. follow a normal distribution $X_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$,
- Case 3: the r.v. follow a Beta distribution $X_i \stackrel{i.i.d.}{\sim} \beta(1, 2)$.

The found GEV parameters, resulting from the maximum log-likelihood estimation, with their 95% confidence interval are given in the Table below.

	$\hat{\xi}$	$\pm \delta\hat{\xi}$	$\hat{\mu}$	$\pm \delta\hat{\mu}$	$\hat{\sigma}$	$\pm \delta\hat{\sigma}$
Case 1	0.17	0.11	9.63	1.92	2.03	0.22
Case 2	-0.08	-0.11	7.58	0.36	0.38	-0.03
Case 3	-0.5	-0.53	0.98	0.05	0.005	-0.46

As announced in Sec. 4.3.1, one can see that the first case is a positive shape ($\hat{\xi} \pm \delta\hat{\xi} > 0$) corresponding to a Fréchet distribution. The second case is a shape around zero ($-0.19 \leq \hat{\xi} \leq 0.04$) and corresponds to a Gumbel distribution. The last case, with $-1.02 \leq \hat{\xi} \leq 0.03$, can be associated to a Weibull distribution.

Fig. 4.3 below illustrates the diagnostic plots associated to these three cases. The first row shows the empirical distribution of the maxima obtained through the MC simulations (histogram) *vs* the GEV pdf given in (4.15) evaluated with $\hat{\boldsymbol{\theta}}_{GEV}$. The second row shows the Q-Q plots. Each data point in blue defines a sample point and the red straight lines show the diagonal. One can observe, for each case, the linearity of the curves with no significant deviations in the tails validating the estimated GEV model. The last row shows the return level plots with theoretical value (blue) and empirical one (red) with the 95% confidence intervals (black dashed curves, cf. Appendix. D).

One can see the different shapes of the return level plots according to the value of ξ and the concordance between the theoretical plot (red) and empirical one (blue). This confirms again the good evaluation of the fitted GEV model.

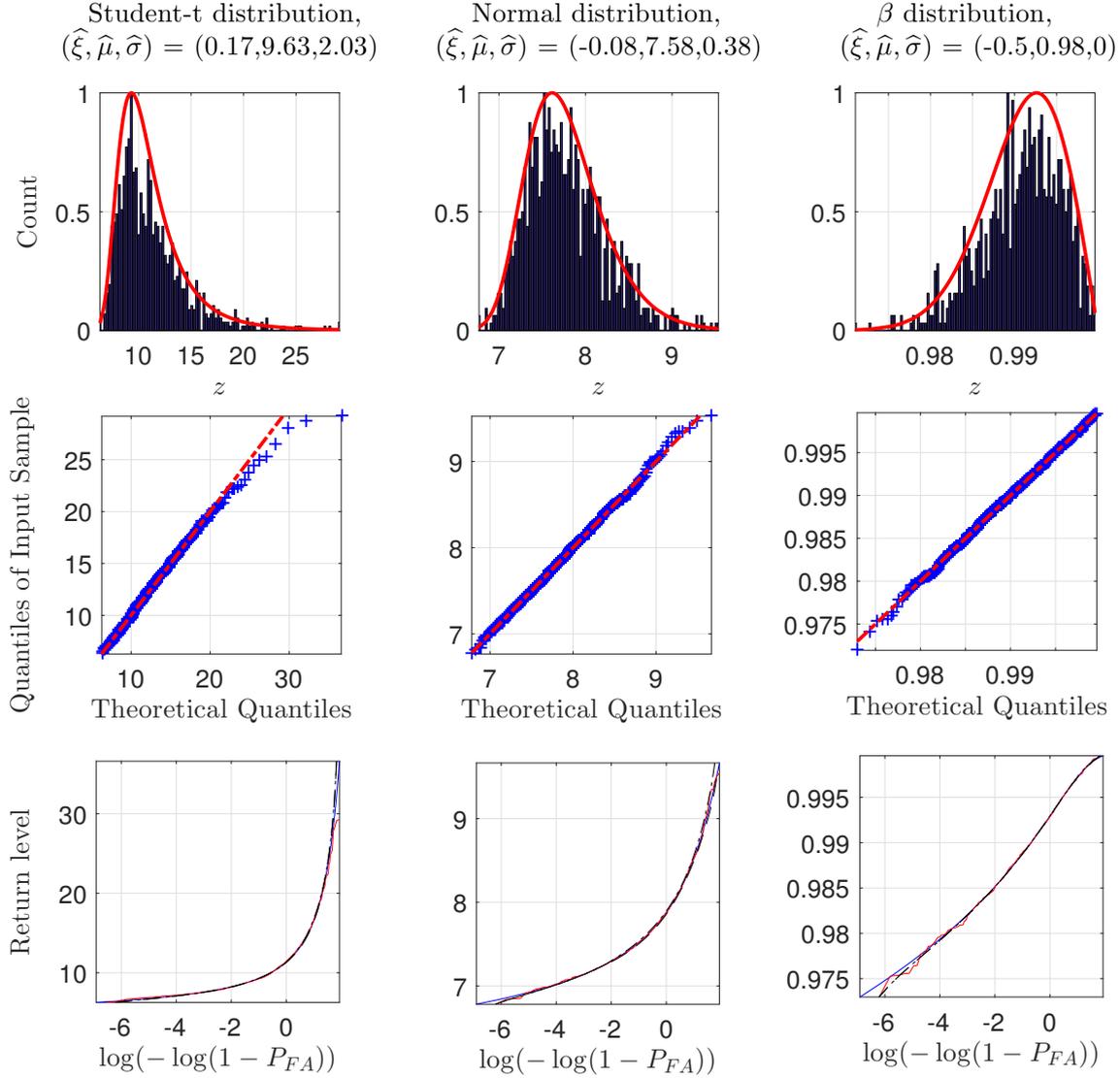


Figure 4.3 – Diagnostic plots for three different GEV estimations.

4.4 False alarm rates evaluation

Let us now describe common methods that we can find in the literature to evaluate the false alarm rate of T_M applied to the LS periodogram. In Chap. 3, we have seen that T_M is easy to set up and appeared powerful to detect very sparse signals in the Fourier space.

4.4.1 Method of “independent” frequencies

We are interested here in the statistical behavior of $M := T_M(\mathbf{P}_{LS})$ with:

$$T_M(\mathbf{P}_{LS}) := \max\{P_{LS}(\nu_1), P_{LS}(\nu_2), \dots, P_{LS}(\nu_N)\} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \quad (4.18)$$

with $P_{LS}(\nu_i)$ a set of N r.v. with distribution $\varphi_{P_{LS}}$ given in (4.8) in the case of a WGN (note that the value of N is here different from the total length of the time series and corresponds to the number of considered Fourier frequencies). For different periodogram standardizations (see,

[Baluev, 2008]) the distribution φ_M changes depending on the number of degrees of freedom used in the orthogonalization (2 for the P_{LS}).

If the N considered ordinates are independent, the distribution of (4.18) is easily derivable:

$$\Phi_M(m) := \Pr(M \leq m) = \prod_{i=1}^N \Pr(P_{LS}(\nu_i) \leq m) = \Phi_{P_{LS}}^N. \quad (4.19)$$

However, in practice the ordinates $P_{LS}(\nu_k)$ are not independent if the sampling is irregular. Consequently, Φ_M never equals $\Phi_{P_{LS}}^N$ in (4.19). Indeed, the P_{LS} ensures the same marginal distribution in case of WGN for the P_{LS} ordinates but dependencies between these ordinates exist because of the irregular sampling. The correlations and level of dependency of the r.v. involved in (4.3) are contained in the spectral window (cf. discussion in [Frescura et al., 2008]).

Common methods consist in considering (4.19) anyway and finding, by empirical methods, an estimate of an “effective” number of “independent” periodogram ordinates, say \hat{N} , and evaluate the P_{FA} as

$$P_{FA}(\gamma, M) \approx 1 - \Phi_{P_{LS}}^{\hat{N}}.$$

Examples of empirical methods to estimate \hat{N} are:

- Methods using MC simulations [Horne and Baliunas, 1986],
- Method deriving analytical estimates through different assumptions [Schwarzenberg-Czerny, 1998, Cumming, 2004, Frescura et al., 2008]. For example, Schwarzenberg-Czerny uses the approximation that the distribution of the maximum (well normalized) periodogram component is related to the distribution of a single component.
- Methods using bootstrap procedures (cf. paragraph below) to create thousands of realizations of the considered test statistics. \hat{N} is then estimated through the median value of the obtained empirical maxima (say m_{med}) as $\hat{N} = \frac{\log 0.5}{\log \Phi_M(m_{med})}$ [Schwarzenberg-Czerny, 2012].

However, in general, these empirical methods lead to inconsistencies because the supposed functional form of the P_{FA} above is wrong [Frescura et al., 2008, Süveges, 2014]. In the absence of reliable estimates for the distribution of test statistics based on P_{LS} (or possibly an other periodograms of unevenly sampled time series), one can turn to pure bootstrap techniques.

4.4.2 Bootstrap methods

In essence, bootstrap procedures use sample-driven Monte Carlo simulations to produce empirical estimates of the distribution of a considered set of random variables. These techniques, initially introduced for i.i.d. r.v., have been intensively studied in the last decades (see e.g. [Zoubir and Iskander, 2004]). Bootstrap techniques exist in the case of even sampling and weakly dependent data (e.g. [Paparoditis and Politis, 1999]) and different resampling procedures have been proposed (e.g. [Politis and Romano, 1994]). For longer memory processes, “AR-aided” periodogram bootstrap consists in estimating the parameters of an AR model in the time domain and in applying a non-parametric kernel based correction in the frequency domain to counteract the effects of estimation noise on the AR parameters (e.g. [Kreiss and Paparoditis, 2003]). For the uneven sampling case, “dependent wild bootstrap” adapting to weakly dependent data unevenly sampled can be found in [Shao, 2010] and a bootstrap procedure using a Generalized Extreme Value estimate on the periodogram maxima is

presented in [Süveges, 2014] (cf. Sec. 4.4.3).

As underlined in [Young, 1994] for instance, the choice of the parameters of such procedures (e.g. parametric noise models, number of blocks, block lengths, data weights or repartition of the sampled blocks) can be delicate. Moreover, the use of parametric methods like AR modelling is obviously not well adapted to irregular sampling (since the AR model involves values which are potentially missing, one have to reconstruct the missing data cf e.g. [de Waele and Broersen, 2000, Bos et al., 2002, Baghi et al., 2016]).

For irregular sampling, in astronomy, one can find two classical families of bootstrap procedures to derive the empirical distribution of the test statistics:

1. The first family of bootstrap consists in generating WGN with same variance as the observed (potentially CLEANed) data, sampling them as the observation, and computing the test statistics (generally, T_M) on each obtained periodogram to evaluate its distribution.
2. The second family consists in randomly shuffling the observation (keeping the time sampling epochs fixed), computing the test statistics on each obtained periodogram, and evaluating its distribution.

4.4.3 FA evaluation using theory of the extremes

Recent approaches proposed to use parametric models for the periodogram maxima. For example, Baluev proposed a method based on the extremes of continuous-parameter stochastic processes [Baluev, 2008]. Considering different distributions (in particular the beta, F or χ^2 distributions), Baluev derived an analytical upper-limit for the P_{FA} of $T_M(\mathbf{P}_{LS})$. This upper-limit was derived assuming small aliases and weak spectral leakage in the observed time series and has been numerically proved to be efficient for small P_{FA} ($< 5\%$) but not for higher ones. This can be problematic when several signals are present in the data [Süveges et al., 2015].

Other interesting extrema distributions are the GEV, which are the limiting distribution of maxima of sequence of possibly dependent r.v. (cf. Sec. 4.3). In practice, the GEV parameters estimation is usually done on data blocks as only one observed time series is available. The choice of the blocks' size is critical and has to be done carefully as, on the one hand, large blocks leads to few maxima and consequently to large variance in the GEV estimation. On the other hand, small blocks could invalid the assumption related to independence between the z_k components in Theorem. 3 and lead to biases.

Süveges proposes a procedure combining a GEV model for the periodogram maximum and a bootstrap procedure to evaluate the P_{FA} associated to test T_M in the case of unevenly sampled time series involving WGN [Süveges, 2014, Süveges et al., 2015]. In few words, the procedure consist in the following steps:

1. Evaluate the LS periodogram of the original time series using (4.3),
2. Make a number (say R) of non parametric bootstraps of this time series (cf. second family of bootstrap described in Sec. 4.4.2),
3. Take the set of maxima $\{m_j\}$, $j = 1, \dots, R$, of the R resulting partial periodograms,
4. Estimate a $\hat{G}(\mathbf{m}, \hat{\boldsymbol{\theta}}_{GEV})$ model on these partial maxima (using (4.15) in Sec. 4.3.2),
5. Extrapolate for the periodogram of the original time series, i.e. use $\hat{G}(\mathbf{m}, \hat{\boldsymbol{\theta}}_{GEV})$ in Eq. (4.16) to find the threshold of the target P_{FA} ,
6. Estimates the confidence interval (cf. Appendix. D).

This hybrid method allows to derive reliable $\gamma(\widehat{P}_{FA})$ for all target false alarm rates, and allows to derive confidence intervals of the estimated return levels when the noise is uncorrelated. Moreover, this procedure allows to decrease the number of MC used in the classical bootstrap procedures, as described in Sec. 4.4.2, and hence allows to save a significant amount of computational time.

However, this technique is not directly applicable when considering colored noise as the shuffling of the data, made during the bootstrap, breaks the intrinsic data correlations and consequently affects the derived empirical test statistics distribution. In this spirit, we will attempt to take benefit from this bootstrap/GEV technique when applied to standardized periodogram of the form of (3.2). This point is developed in Sec. 4.5.2, which extends under some working assumptions the Süveges' procedure to the case of irregularly sampled time series involving colored noise.

4.4.4 Conclusions

To conclude, irregular sampling creates several peaks in the periodogram due to the spectral window that looks like “noise” or “signal”.

The Lomb-Scargle periodogram is equivalent to a least-square analysis and several variants of this periodogram have been proposed (e.g. Generalized LS or Keplerian periodograms).

Deriving the false alarm rate of tests based on such periodograms ordinates is not easy as the different ordinates are not independent. Any procedure based on the Lomb-Scargle periodogram has to face the difficulty that the dependence of some of these ordinates is strong, even in the white noise case (this is well discussed in [Süveges, 2014] for instance). Besides, even when considering the marginal distribution of the ordinates, the convenient exponential distribution (χ_2^2) of Lomb-Scargle is only valid in the white noise case, so the main advantage of this method vanishes for colored noise. So, if such a method is used for partially unknown colored noise, the false alarm rate assuming white Gaussian noise will almost surely be wrong. In fact, it is not clear what would be the benefit of using this procedure in the colored noise framework, but if this is done anyway, specific ways to estimate the resulting false alarm rate remain to be derived.

When the noise is colored, several methods (based on least-square analysis, maximization of likelihood functions or sparse regularization) consist in modelling the noise by a specific parametric model (e.g. Gaussian processes). However, the FA control is difficult as the derived significance is model dependent. In practice, such methods can lead to very different detection significance results when applied to extrasolar planet detection.

To control the FA, one can turn to bootstrap analyses, which consist in estimating the distribution. If the test is based on r.v. extrema, combined with the bootstrap procedure, a GEV model can be used to save a significant amount of computational time. However, controlling reliably the FA rate in sinusoid detection is a problem in the case of an unevenly sampling and unknown colored noise, even when resorting to bootstrap techniques (the Süveges' technique was developed for white Gaussian noise only).

In this study, we renounce to obtain analytical expression of the P_{FA} of tests based on (any) periodogram ordinates. We propose instead, in Sec. 4.5.1, an original and automated bootstrap method allied to periodogram standardization. We will see in Sec. 4.5.2 that the method can take advantage from GEV distributions (as proposed in [Süveges, 2014, Süveges et al., 2015]) to save computational time. The proposed procedure will be then numerically studied on a synthetic colored noise in Sec. 4.3.

4.5 Proposed bootstrap procedure

The present section explores a bootstrap solution combined with independent simulations of the colored noise to solve the problem of controlling the FA rates in case of uneven sampling and colored noise. As in Chap. 3, we assume that a training data set \mathcal{T}_L of the noise is available. However, this set provides only a *small* number of noise samples - typically, the order of the number of data samples. In this regime, our knowledge of the statistics of the random process under \mathcal{H}_0 remains strongly impacted by estimation noise. Our strategy is again to get rid of any nuisance effect from the noise PSD by considering standardized periodograms of the form (3.2), for which the $\{t_j\}$ in (2.24) are unevenly spaced, and to capture the dependencies between the periodogram ordinates through bootstrap techniques.

4.5.1 Direct bootstrap approach

Let us consider an observed time series of PSD S , noted $X_{obs}|S$, involving a random process (with unknown PSD S_E), plus possibly a periodic component, sampled at uneven time instants $\{t_j\}_{j=1,\dots,N_{obs}}$. We perform the test $T_M(\mathbf{Z}) := Z_{(N)}$, with \mathbf{Z} some periodogram, to decide between the presence or absence of the periodic component. The question is to evaluate the resulting P_{FA} .

The study is presented on the classical periodogram, but any other periodograms described in Sec. 4.1 could also be used without loss of generality. We also assume here that all epochs are multiple of a common, possibly vanishingly, small interval δt .

Let us describe the considered test more precisely. First, we compute the periodogram of the data. To reduce the computational time, we consider here the classical periodogram (2.24):

$$P(\nu_k; X_{obs}|S) = \frac{1}{N} \left| \sum_{j=1}^N X_{obs}(t_j) e^{-i2\pi\nu_j\delta t} \right|^2.$$

as it does not significantly differ from the LS periodogram and is easier to compute. Others periodograms can, obviously, be used as the distribution of statistical tests based on any periodogram is not known when the sampling is irregular.

Second, we compute the averaged periodogram (2.48) from the L time series of the training data set, sampled at the same time instants $\{t_j\}$, which we note $\{X_\ell|S_E\}_{\ell=1,\dots,L}$:

$$\bar{P}_L(\nu_k; \{X_\ell|S_E\}) := \frac{1}{L} \sum_{\ell=1}^L \frac{1}{N} \left| \sum_{j=1}^N X_\ell(t_j) e^{-i2\pi\nu_j\delta t} \right|^2.$$

Specifically, the ordinates of the standardized periodogram in (3.2) are obtained by:

$$\tilde{P}(\nu_k; X_{obs}|S_E, \{X_\ell|S_E\}) := \frac{P(\nu_k; X_{obs}|S)}{\bar{P}_L(\nu_k; \{X_\ell|S_E\})}. \quad (4.20)$$

We run test T_M on the vector of standardized ordinates $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$ (4.20):

$$T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L) := \max_k \frac{P(\nu_k; X_{obs}|S)}{\bar{P}_L(\nu_k; \{X_\ell|S_E\})}. \quad (4.21)$$

In the following, we call M the maximum $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ to be consistent with Sec. 4.4.1. The associated P_{FA} is then defined as:

$$P_{FA}(M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L), \gamma) := \Pr(M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L) > \gamma | \mathcal{H}_0). \quad (4.22)$$

Our goal is to obtain a confidence interval for this P_{FA} at any γ .

Of course, if we could generate as many realizations as desired of the r.v. $M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ (resulting in, say b realizations $\{m_i\}_{i=1,\dots,b}$ of the r.v. M , collected in a vector $\mathbf{m} := [m_1 \dots m_b]^\top$), we would be able to estimate the empirical distribution of the maximum and thus the P_{FA} by:

$$\hat{P}_{FA}(\gamma; \mathbf{m}) := 1 - \hat{\Phi}_M(\gamma; \mathbf{m}), \quad (4.23)$$

with $\hat{\Phi}_M$ the empirical cdf of M , and where the dependency of the estimates on the observed values \mathbf{m} is explicit. The principle of such P_{FA} estimation procedure is illustrated in Fig. 4.4.

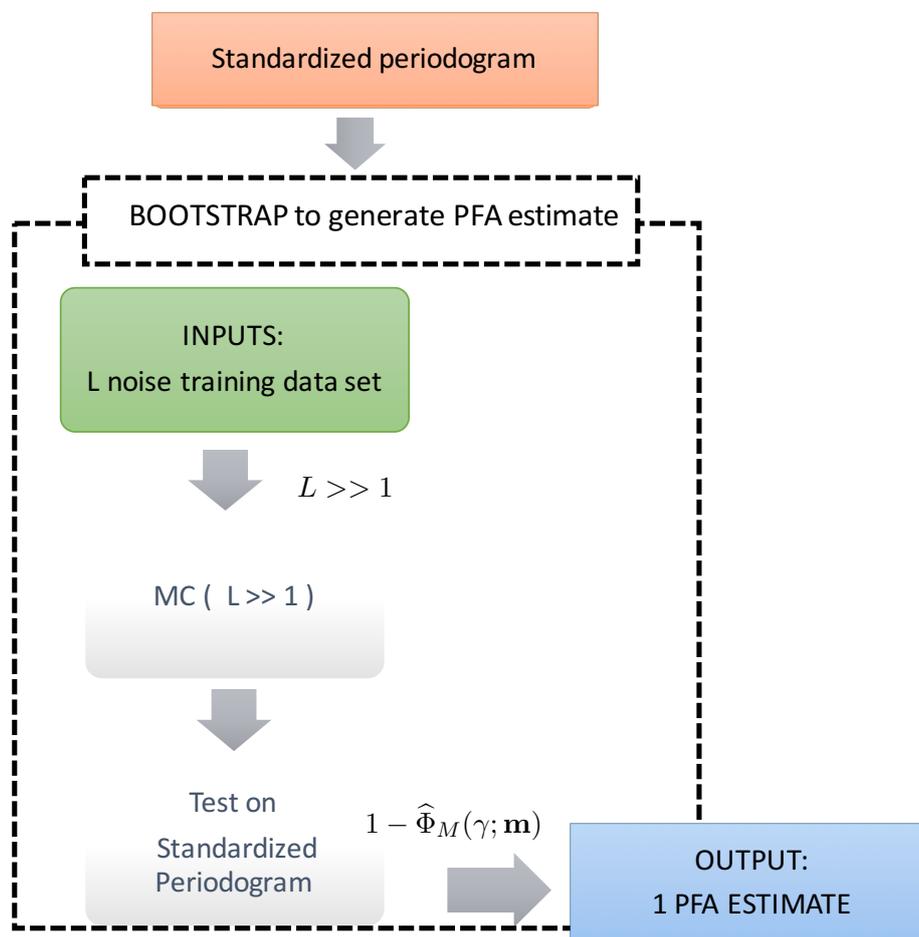


Figure 4.4 – P_{FA} estimation procedure (possible only if $L \gg 1$).

We can not do so however, because the training data set has finite size L . To counteract this fact, we may estimate the parameters of the noise from $\{X_\ell|S_E\}_{\ell=1,\dots,L}$ under some model, with the aim of generating secondary noise data sets according to this estimated model. To make an example, we opt for a general parametric colored noise model as an autoregressive process with PSD given in (2.47). As seen in Sec. 3.4, various selection criteria exist to estimate the order of the process. Examples of such criteria are given in (3.43) and depend on the estimated AR parameters $\hat{\theta}_{AR} := [\hat{\sigma}, \hat{c}_j, \hat{\sigma}_\sigma^2]^\top$, from which a PSD estimate \hat{S}_E can be obtained as in (2.47) (we have seen in Chap. 3 - cf. Sec. 3.5.5 - that this estimate is always noisy; this estimation noise will however be accounted for by the proposed bootstrap procedure). For example, in the simulations of Sec. 4.6, we will use the Bridge criterion (G) defined in (3.43) to estimate S_E . The estimated AR parameters also allow to generate $(L + 1) \times b$ correlated time series, noted $\{X_\ell|\hat{S}_E\}$, allowing to generate simulated standardized periodograms $\{\tilde{P}(\nu_k; X|\hat{S}_E, \{X_\ell|\hat{S}_E\})\}$, each requiring 1 time series for numerator (2.24) and L for denominator (2.48). The resulting b maxima can be used to estimate the P_{FA} with (4.23). The principle of such P_{FA} estimation procedure is illustrated in Fig. 4.5.

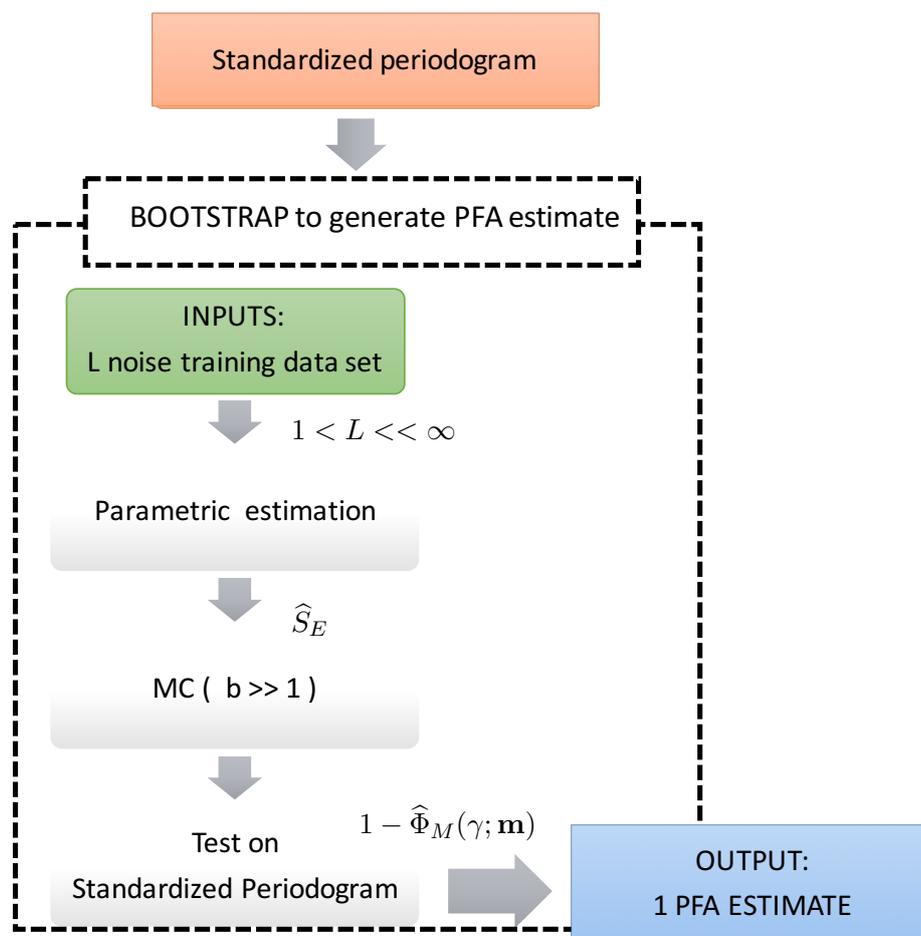


Figure 4.5 – P_{FA} estimation procedure if $1 < L \ll \infty$.

However, as just recalled above, this estimate is obtained by means of *one* set of AR parameters estimated from \mathcal{T}_L and is thus random. Obviously, we would have obtained a different estimate for a different training set \mathcal{T}_L . The question is therefore to evaluate the distribution of the P_{FA} estimate. We can not evaluate this distribution (called $\mathcal{D}_{\hat{P}_{FA}}$ below) because it requires a very large number of training data sets L . This is where the bootstrap really comes in: we propose to use the estimated parametric model parameters (e.g. the AR coefficients $\{\hat{c}_j\}$) on the time series of \mathcal{T}_L to generate a number B of “fake training data sets”, from which we can obtain an estimate of distribution $\mathcal{D}_{\hat{P}_{FA}}$, called $\hat{\mathcal{D}}_{\hat{P}_{FA}}$ below. This leads to the bootstrap procedure, called B_0 , which follows the five main steps above, described in details in Algorithm. 1 and illustrated in Fig. 4.6.

- (i) **1st PSD estimation** (2.47). It begins by a first parametric estimation on \mathcal{T}_L using (3.43) (row 2).
- (ii) **2nd PSD estimation**. Secondary PSD estimates (noted $\hat{S}_E^{(i)}$, $i = 1, \dots, B$) are generated by estimating (a second time) the parametric noise model parameters on a i -th fake training data set noted $\{X_\ell | \hat{S}_E\}$ using (3.43) (for simplicity, we omit the dependence in i in this notation, row 4).
- (iii) **P_{FA} estimate**. These parameters are used to generate b (loop on j , row 5) standardized periodograms using (4.20), the corresponding maxima $\mathbf{m}^{(i)}$ using (4.21) and P_{FA} estimates using (4.23) with $\mathbf{m}^{(i)}$ (row 6).
- (iv) Repeat (ii) - (iii) to generate a set of B P_{FA} estimates (loop on i , row 3).
- (v) The distribution $\hat{\mathcal{D}}_{\hat{P}_{FA}}$ is then estimated from the set of B such estimates, $\{\hat{P}_{FA}^{(i)}\}_{i=1, \dots, B}$ (row 10).

Algorithm 1 Proposed bootstrap procedure (B_0)

```

1: procedure  $B_0$  ( $\mathcal{T}_L$ )
2:    $\hat{S}_E(\nu_k; \hat{\theta}_{AR}(\{X_\ell | S_E\}))$ 
3:   for  $i = 1, \dots, B$  do
4:      $\hat{S}_E^{(i)}(\nu_k; \hat{\theta}_{AR}(\{X_\ell | \hat{S}_E\}))$ 
5:     for  $j = 1, \dots, b$  do
6:        $m_j^{(i)} = \max_k \tilde{P}(\nu_k; X | \hat{S}_E^{(i)}, \{X_\ell | \hat{S}_E^{(i)}\})$ 
7:     end for
8:      $\hat{P}_{FA}^{(i)}(\gamma; \mathbf{m}^{(i)})$ 
9:   end for
10:  return  $\hat{\mathcal{D}}_{\hat{P}_{FA}}(\{\hat{P}_{FA}^{(i)}\})$ 
11: end procedure

```

One advantage of this procedure is that it is independent from the data under test. We note, however, that the choice of the parametric model (e.g., AR, MA, ARMA, ARIMA, Harvey functions or power functions) has to be done cautiously, and can be verified using diagnostic plots.

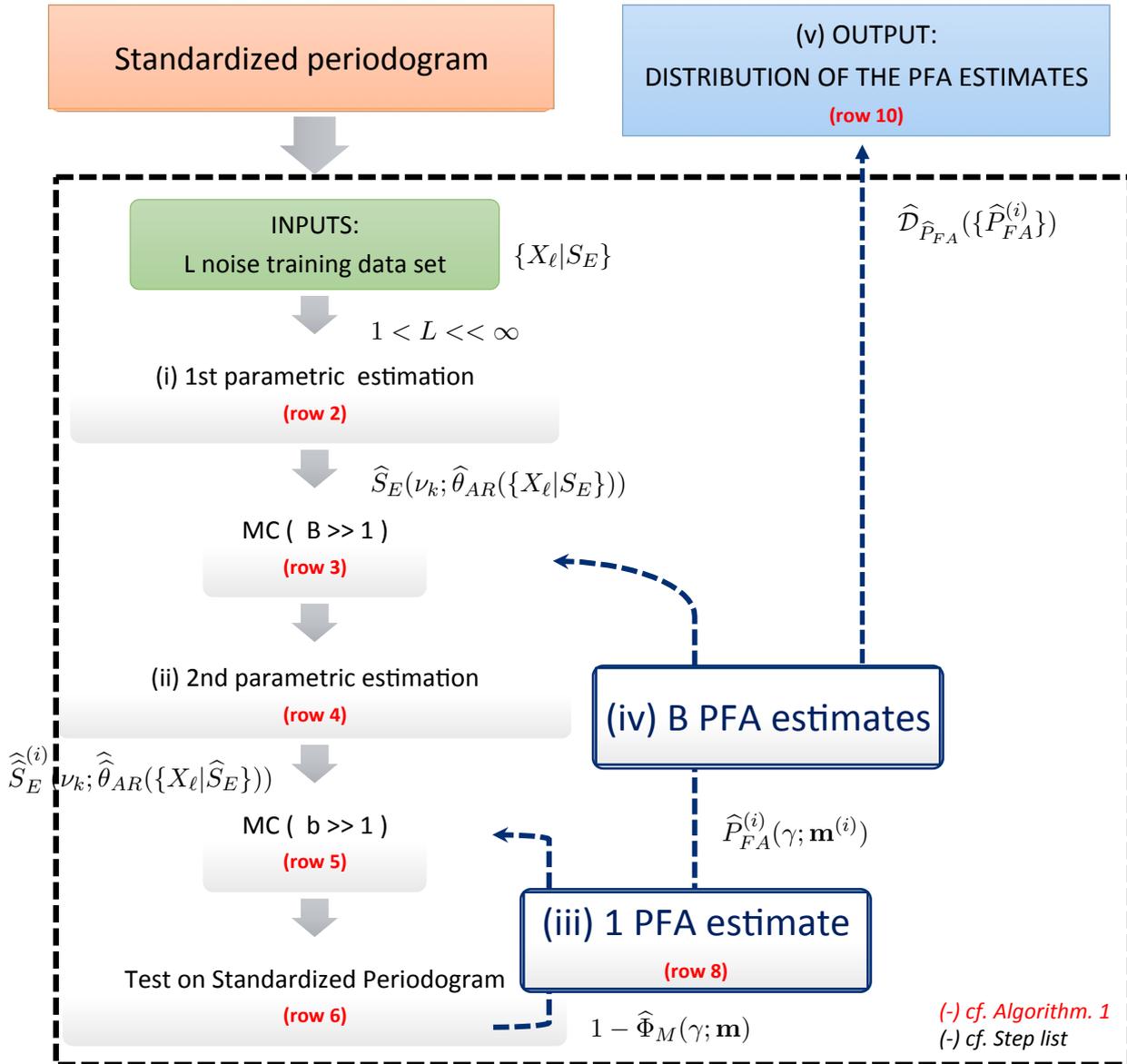


Figure 4.6 – Estimate of the P_{FA} estimates distribution derived by B_0 procedure. The different rows of Algorithm. 1 are indicated in red, and the main steps in black.

4.5.2 “Accelerated” bootstrap approach

The B_0 procedure above does not rely on any model for the cdf of M , $\widehat{\Phi}_M(\gamma; \mathbf{m}^{(i)})$, in (4.23). To be efficient, this estimation requires b to be large, which makes it computationally expensive. However, as presented in Sec. 4.3, interesting results from univariate extreme-value theory show that the maximum of a set of identically distributed (dependent or not) r.v. follows a GEV distribution [Coles, 2001]. This suggests that GEV distributions can be used as a model for the cdf of $\mathbf{m}^{(i)}$. This method was used in [Süveges, 2014] in the case of white noise but we show below that it can also be used in the case of colored noise when a training data set is available, using the considered standardization.

The GEV model impacts only step 8 of Algorithm. 1, and the other steps are unchanged. In Algorithm. 1, the unknown parameters $\boldsymbol{\theta}_{GEV} = [\xi, \mu, \sigma]^\top$ can be estimated using $\mathbf{m}^{(i)}$ (step 8), for instance by maximum likelihood as described in Sec. 4.3.2 (cf. (4.17)). In this case the maximization relies on the iterative method described in [Coles, 2001] and not reproduced here. Denoting by $\widehat{\xi}^{(i)}$, $\widehat{\mu}^{(i)}$, and $\widehat{\sigma}^{(i)}$ the estimated GEV parameters, $\widehat{P}_{FA}^{(i)}$ can be estimated using (4.15) by

$$\widehat{P}_{FA}^{(i)}(\gamma; \mathbf{m}^{(i)}) := 1 - \widehat{G}(\gamma; \widehat{\boldsymbol{\theta}}_{GEV}), \quad (4.24)$$

In the following, we will note by B^* the B_0 procedure using the GEV model.

4.6 Numerical studies on a synthetic noise process

The present numerical study only involves time series under \mathcal{H}_0 as we are interested in controlling the false alarm rates. The results are obtained using a synthetic colored noise in order to analyze and validate the B_0 procedure and its “GEV-based” form B^* . The different objectives of the following subsections are:

- To illustrate the proposed procedure,
- To validate B_0 and B^* for different samplings schemes and AR estimation criteria,
- To compare B_0 and B^* in terms of performance, behaviours, and computation time.

4.6.1 Validation on known processes

Illustration of the procedure

We first consider a synthetic colored noise, E , as the AR(6) process (with coefficients $c = \{0.7, 0.05, 0, 0.3, 0, -0.3\}$) introduced in Sec. 3.4. The corresponding theoretical PSD S_E is given by (2.47). It is represented by the green curve in the panels d) and e) of Fig. 4.7.

Fig. 4.7 and Fig. 4.8 illustrate a snapshot of some steps and quantities involved in B_0 (see Algorithm. 1). For the uneven sampling, we consider a regular grid $\{t_k := k \Delta t\}_{k=1, \dots, N}$ with $N = 1024$ and $\Delta t = 1$ (panel a), yellow dots), from which we randomly select $N = 103$ data points (black dots). The spectral window of such sampling is illustrated in b), and the periodogram (computed for the yellow dots of panel a)) is shown in panel c).

For \mathcal{T}_L , we generate $L = 20$ synthetic time series following the true PSD S_E , noted $\{X_\ell | S_E\}$, from which we can compute an averaged periodogram $\overline{P}_L | S_E$ (panel d), black solid line). The dashed red curve in panel d) illustrates the primary parametric PSD estimate \widehat{S}_E obtained from $\{X_\ell | S_E\}$ (row 2 of Algorithm. 1). In panel e), the blue dashed line illustrates one secondary PSD estimate $\widehat{S}_E^{(i)}$ obtained from a “fake training data set” $\{X_\ell | \widehat{S}_E\}$ of L simulated time series (row 4 of Algorithm. 1). This panel also shows for comparison with d) the averaged periodogram $\overline{P}_L | \widehat{S}_E$ (red solid line). Panel f) displays one standardized periodogram obtained

from $L + 1$ series $\{X_\ell|\widehat{S}_E^{(i)}\}$, with the maximum value $m_j^{(i)}$ indicated by the red circle (row 6 of Algorithm. 1).

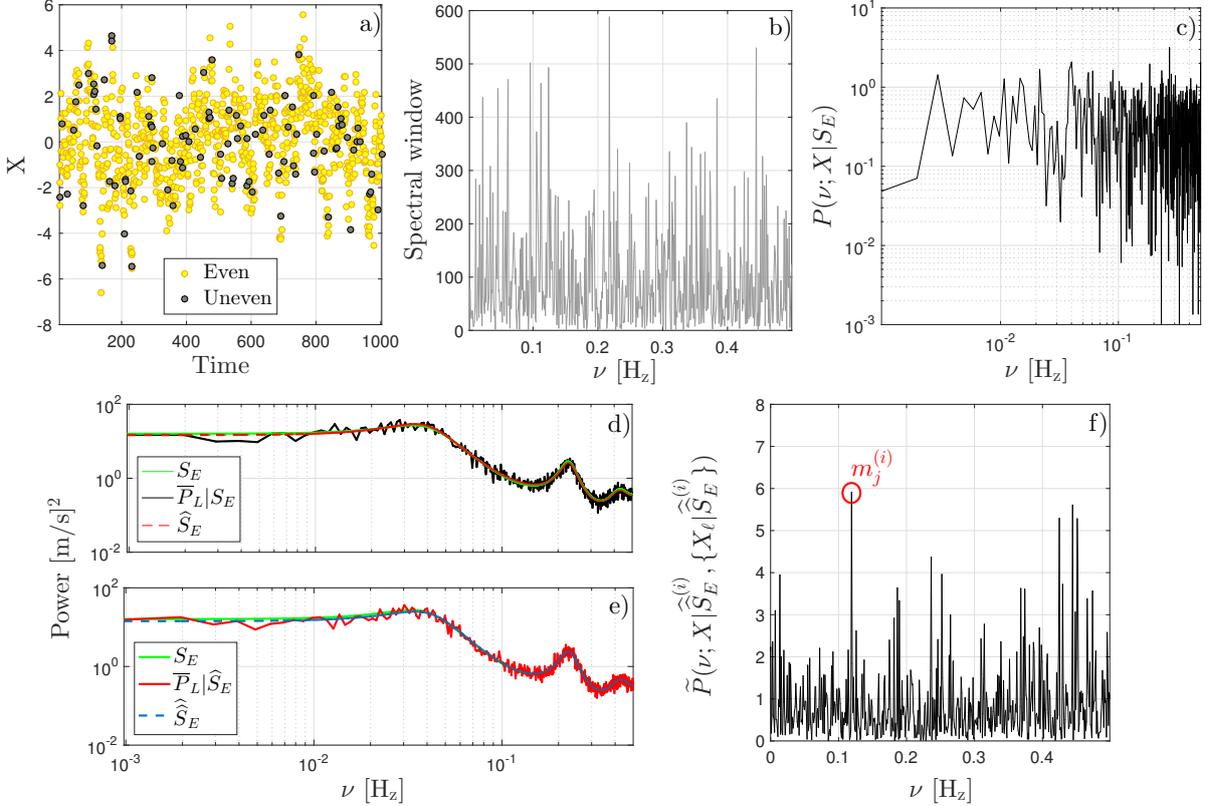


Figure 4.7 – Snapshots of some steps in the B_0 procedure for an AR(6) colored noise process.

The performance in terms of P_{FA} estimates of the B_0 procedure are shown in Fig. 4.8. In all the following figures (i.e. until the end of this chapter), the green curves will represent the true P_{FA} given in (4.22) (more precisely, this is an accurate estimation of the true P_{FA} obtained through 10^5 MC simulations of the investigated AR process. It will serve as reference for comparing the performance of the proposed procedure).

The left panel of Fig. 4.8 shows the B ($= 5200$, here) estimates

$$\widehat{P}_{FA}^{(i)}(\gamma) = 1 - \widehat{\Phi}_M^{(i)}(\gamma; \mathbf{m}^{(i)}), \quad \text{for } i = 1, \dots, B,$$

obtained through the B_0 procedure with $b = 10^4$ (grey cruves).

The right panel shows, in grey, the distribution $\widehat{\mathcal{D}}_{\widehat{P}_{FA}}(B_0)$ for a threshold fixed at $\gamma_0 = 10.6$ corresponding to a true P_{FA} of 0.1 (green). We see that the estimates bound the true value with a relatively small dispersion. A 95% confidence interval, obtained from a Gaussian approximation (red), is indicated in blue and indeed contains the true P_{FA} . A similarly reliable estimation of the true P_{FA} is possible over the whole γ range.

Method ignoring the noise correlations

In Fig. 4.9, we illustrate the P_{FA} estimates obtained when using a classical bootstrap procedure, that would ignore noise correlations. We use the Generalized Lomb Scargle periodogram (P_{GLS}) [Zechmeister and Kürster, 2009], which is a periodogram frequently used in Astronomy. In this procedure, we generate unevenly sampled light curves by resampling the data (through random

permutations), evaluate the P_{GLS} , compute the maximum and perform test (2.51) on

$$\tilde{P}_{GLS}(\nu) := \frac{P_{GLS}(\nu)}{\hat{\sigma}_W^2}$$

with $\hat{\sigma}_W^2$ the variance estimated from the data. This provides one estimate $\hat{P}_{FA}(\gamma)$ and we repeat this experiment 10^3 times. The resulting curves $\hat{P}_{FA}(\gamma)$ are plotted in black. The true P_{FA} of this procedure (in green) has been generated with 10^5 MC simulations. The method fails to estimate accurately the FA rate, because this simple resampling by permutation breaks the data correlation. This is in clear contrast with the procedure shown in Fig. 4.8, which takes benefit from the training data set.

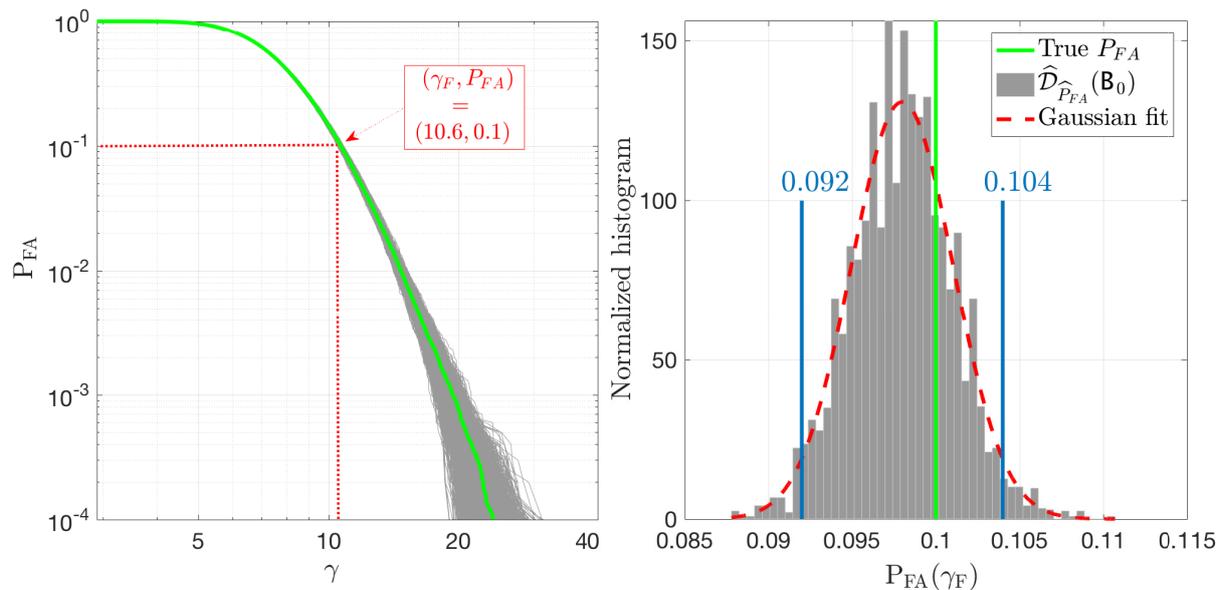


Figure 4.8 – Left: Example of \hat{P}_{FA} (grey) vs P_{FA} (green) for the proposed B_0 procedure. Right: $\hat{D}_{\hat{P}_{FA}}(B_0)$ for a threshold fixed at $\gamma_0 = 10.6$ corresponding to a true P_{FA} of 0.1 (green).

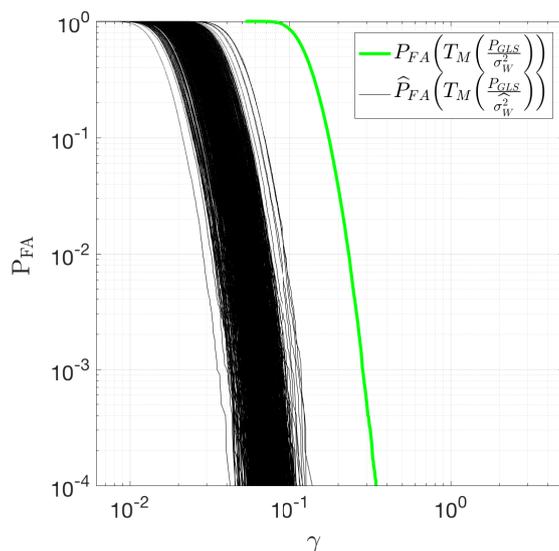


Figure 4.9 – Example of $P_{FA}(\gamma)$ curves obtained using a classical bootstrap procedure ignoring the noise correlations. The true FA is shown in green.

Influence of the data sampling

Let us now investigate the validity of the proposed approach on three other samplings:

- \mathcal{S}_1 : regular sampling ($N_1 = 1024$ points).
- \mathcal{S}_2 : regular sampling with “gaps” ($N_2 = 368$ points = 36% N_1 + 16 gaps).
- \mathcal{S}_3 : random irregular and sparse sampling ($N_3 = 103 = \frac{N_1}{10}$; this case is most representative of real observational in RV data).

The results are shown in Fig. 4.10 to Fig. 4.14 for $b = 10^4$ and $B = 1000$.

- Fig. 4.10 evaluates the distribution of test T_M (4.21) for the three samplings (left to right). The grey histograms are the true distribution found through the true process: they were generated using 10^5 MC simulations of the maximum of periodograms standardized with the true AR process. The different color lines represents the smoothed histograms contours obtained with 10 realizations of \mathbf{B}_0 . One can observe a good match between the histograms showing that the proposed procedure \mathbf{B}_0 allows to reproduce reliable estimates of the test statistics distribution. The histogram modes are stable, around $\gamma = 8$ for the three considered samplings. The distributions obtained using irregular samplings have slightly heavier tails than the one obtained for \mathcal{S}_1 .
- The first row of Fig. 4.11 shows, for the three samplings (left to right), the repartition in function of the frequency of test $T_M(\tilde{\mathbf{P}}|\overline{\mathbf{P}}_L|\mathcal{H}_0)$ in (4.21) for the true process (top row, black). This is the histogram of the frequencies corresponding to the largest peak of the standardized periodogram in (4.20) (c.f. Fig. 4.4). The frequency distribution obtained by the \mathbf{B}_0 procedure (that is, the largest peak in the periodogram obtained for 10 different MC realizations) is shown below (grey). Both distributions seem compatible with a uniform distribution, as desired under \mathcal{H}_0 , and as expected thanks to the periodogram standardization in (4.20).
- Fig. 4.12 illustrates the resulting set of estimates $\{\widehat{P}_{FA}^{(i)}(\gamma)\}$, for $i = 1, \dots, B$ (involved in step. 3 of Algorithm. 1, grey) *vs* the true one (green). One can observe, $\forall \mathcal{S}_i$, $i = 1, 2, 3$, that the estimates are globally distributed around the true $P_{FA}(\gamma)$.
- Fig. 4.13 represents the associated $\widehat{\mathcal{D}}_{\widehat{P}_{FA}}(\mathbf{B}_0)$ for a true $P_{FA} = 90\%$, 50%, 10% and 1% corresponding to the value of γ_0 indicated in each panel. In other words, we set the test $m_j^{(i)} = \max_k \tilde{P}(\nu_k; X|\widehat{\mathcal{S}}_E^{(i)}, \{X_\ell|\widehat{\mathcal{S}}_E^{(i)}\})$ (cf. row 6 of Algorithm 1) to the values γ_0 corresponding to the P_{FA} indicated above, and we investigate the ability of the bootstrap algorithm to identify this P_{FA} . The first row illustrates \mathcal{S}_1 , the second \mathcal{S}_2 , and the last one \mathcal{S}_3 . The empirical distributions of \widehat{P}_{FA} are centred around the true P_{FA} (green) with a small dispersion (the maximum dispersion is observed for $P_{FA} = 0.5$ and is inferior to 0.02 in absolute value).
- Finally, Fig. 4.14 shows the distribution estimates obtained for 10 other realizations of the bootstrap procedure with different irregular samplings \mathcal{S}_3 . One can observe that the reliability of the proposed \mathbf{B}_0 procedure is fairly independent of the considered sampling. One can also note (not shown) that the same statistical results are observed when using the different AR criteria listed in (3.43).

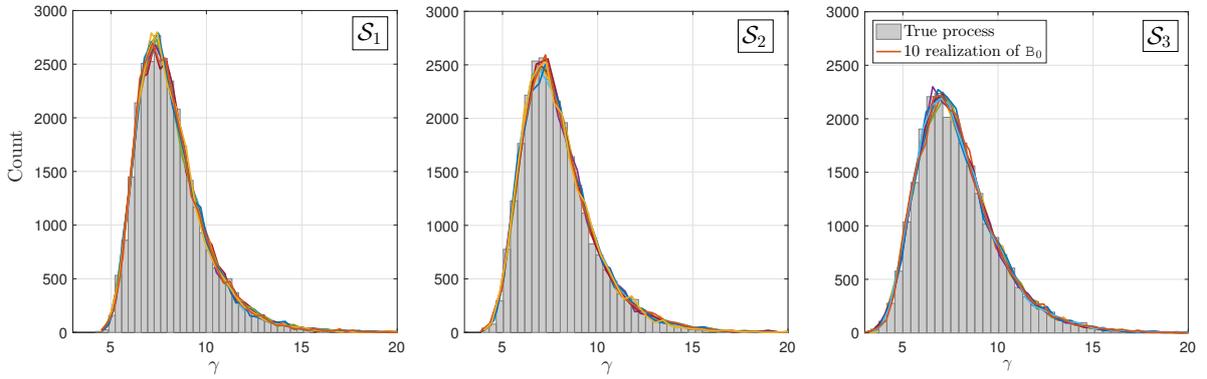


Figure 4.10 – B_0 procedure: Pdf of T_M for the three samplings.

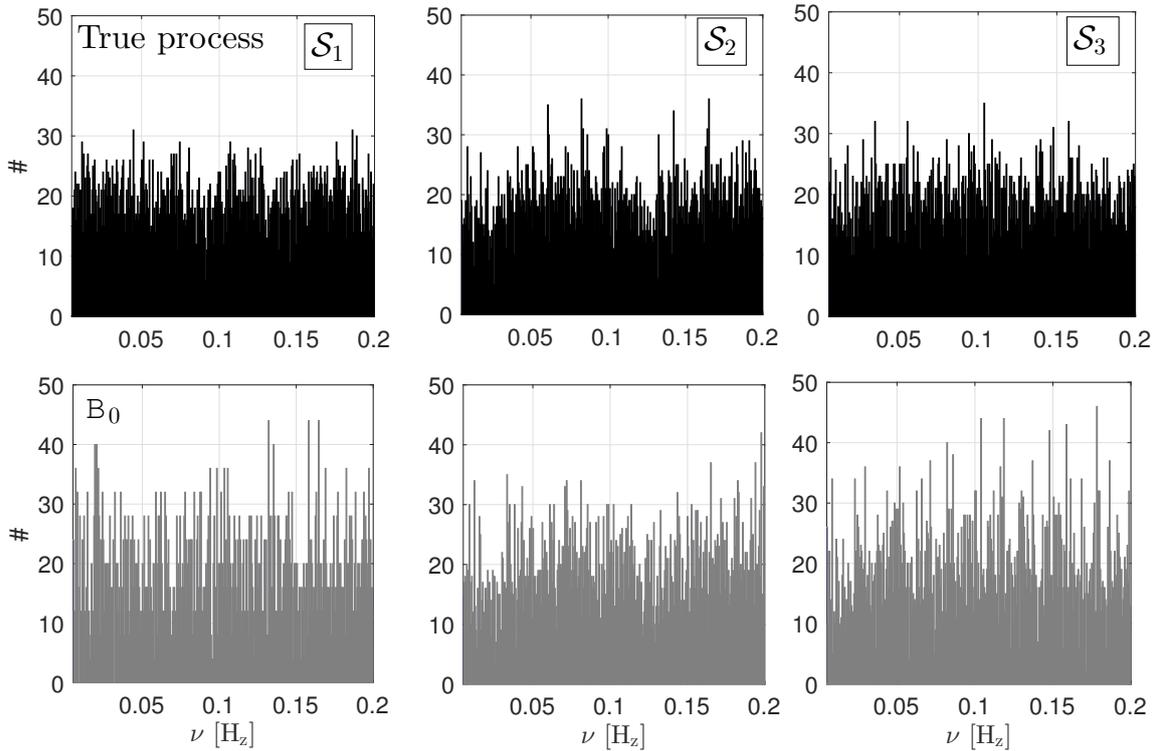


Figure 4.11 – Repartition in frequency of largest periodogram peak identified by test $T_M(\tilde{\mathbf{P}}|\overline{\mathbf{P}}_L|\mathcal{H}_0)$.

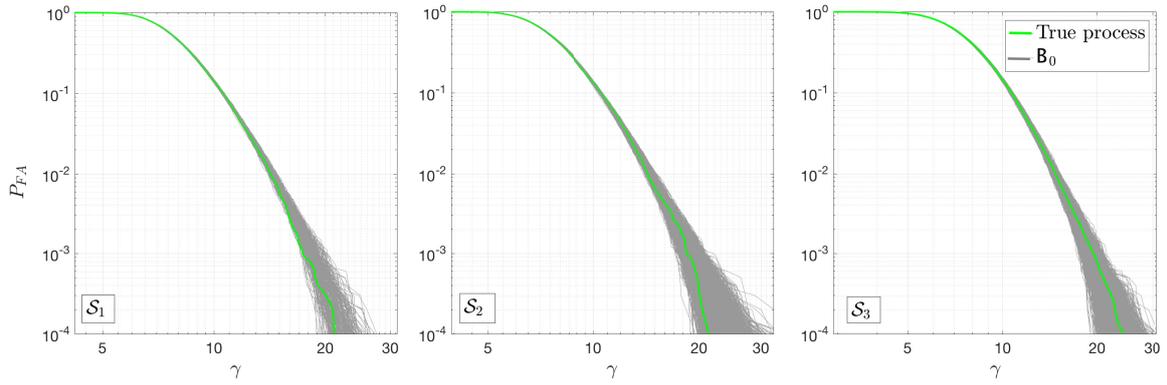


Figure 4.12 – Estimated FA rates: $\hat{P}_{FA}(\gamma)$ vs true $P_{FA}(\gamma)$ for the proposed B_0 for the three considered samplings.

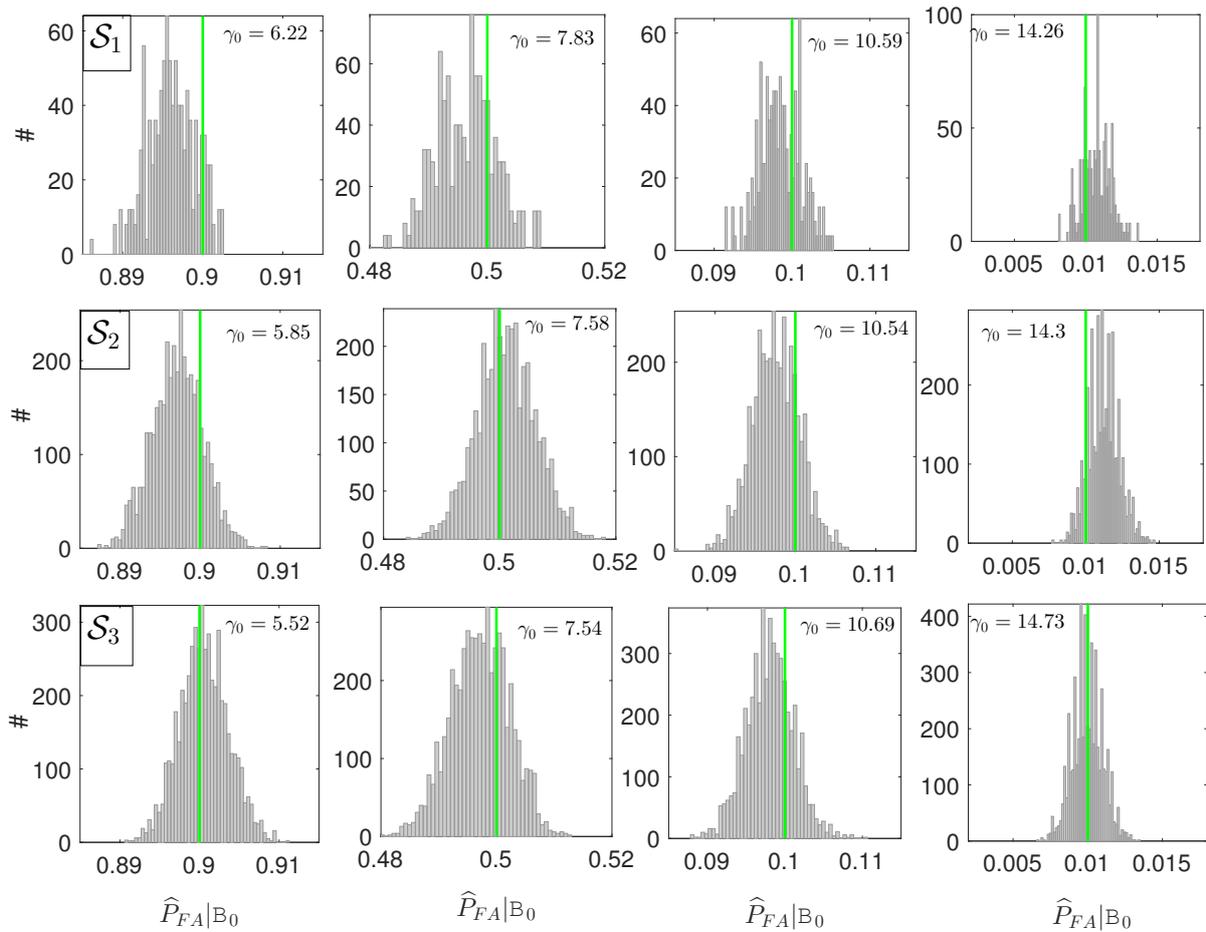


Figure 4.13 – $\hat{\mathcal{D}}_{\hat{P}_{FA}}(B_0)$ for a threshold fixed at γ_0 corresponding to a true P_{FA} (green) of resp. 0.9, 0.5, 0.1 and 0.01 (from left to right). Each row represents one specific sampling.

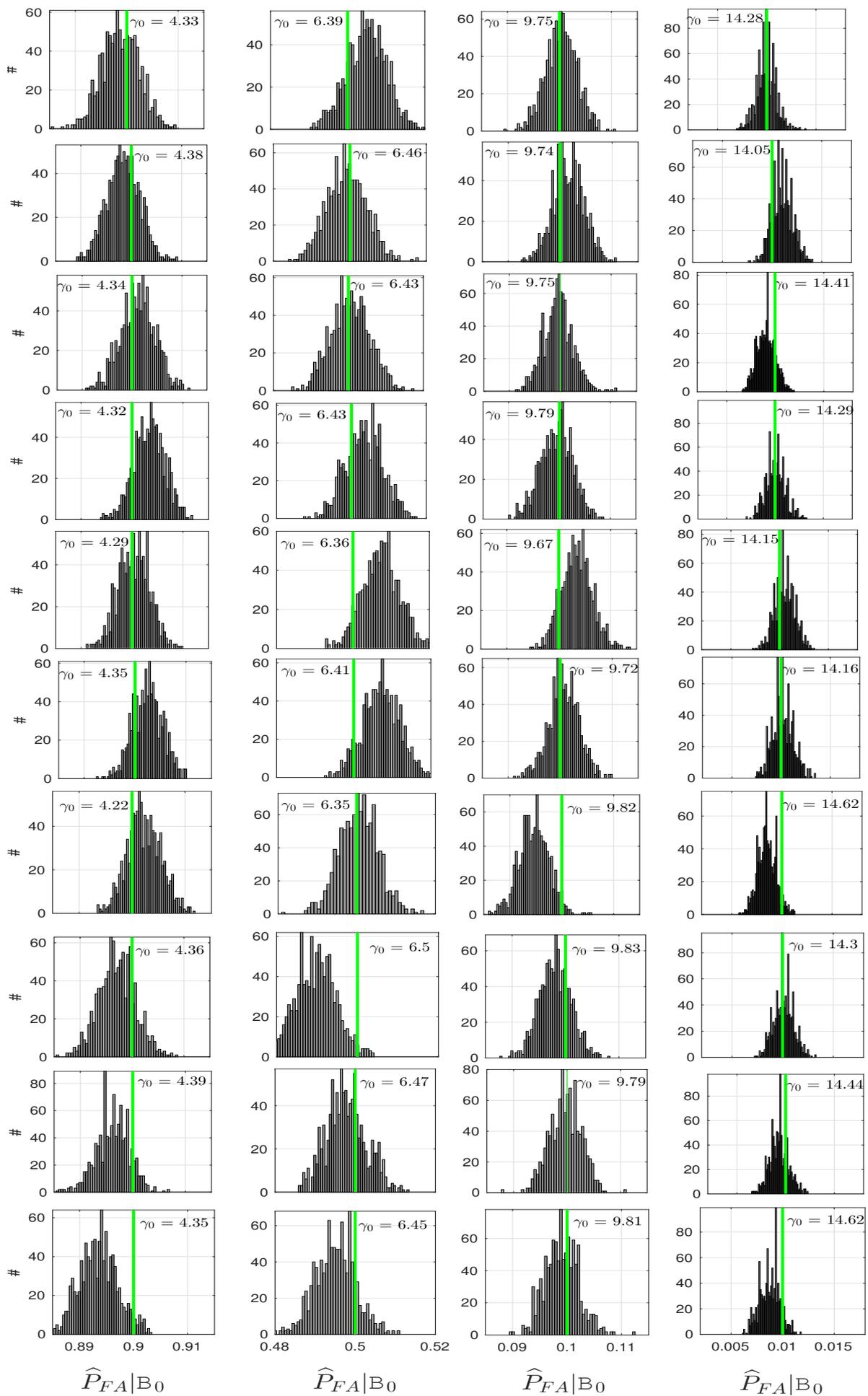


Figure 4.14 $\hat{D}_{\hat{P}_{FA}}(B_0)$ obtained for 10 realizations of the B_0 bootstrap procedure (different rows) with different irregular samplings \mathcal{S}_3 . The true $P_{FA} = 90\%$, 50% , 10% and 1% corresponds to the value of γ_0 indicated in each panel and is shown in green.

4.6.2 “GEV accelerated” algorithm

In practice, one limitation of the B_0 procedure is its computation time, related in particular to the large number (b) of periodogram maxima required for the estimation of the P_{FA} to be accurate. To reduce b while keeping a tight confidence interval for the P_{FA} , we consider applying the version of the bootstrap procedure using the GEV approximation (B^*) as described in Sec. 4.5.2.

False alarm probability: B_0 vs B^* procedures

First, let us compare the reliability of the results based on the GEV model in terms of $\widehat{P}_{FA}(\gamma)$ estimates.

Considering the irregular sampling \mathcal{S}_3 , Fig. 4.15 compares the true P_{FA} (green), the set of estimates $\{\widehat{P}_{FA}^{(i)}\}$ obtained with B_0 (using $B = 5200$ and $b = 10^4$, grey), and the $\{\widehat{P}_{FA}^{(i)}\}$ estimates obtained with B^* (for $B = 1000$, red). Different values of b for B^* are shown (resp. in lexicographical order: $b = 50, 100, 500, 1000$ and $10\,000$). We see an expected behaviour for B^* : the more MC simulations (the larger b) are used during the bootstrap (i.e. 2nd loop, row 5 of Algorithm. 1), the more the GEV approximation gives an accurate estimate of the distribution. For $b = 10^3$ or $b = 500$ (row 2), the distributions get close to the distribution of $\widehat{P}_{FA}(\gamma, B_0)$ computed with 10^4 realizations. For $b = 10^4$ the GEV distribution is more narrow than the distribution obtained with B_0 due to the parametric-aided nature of B^* .

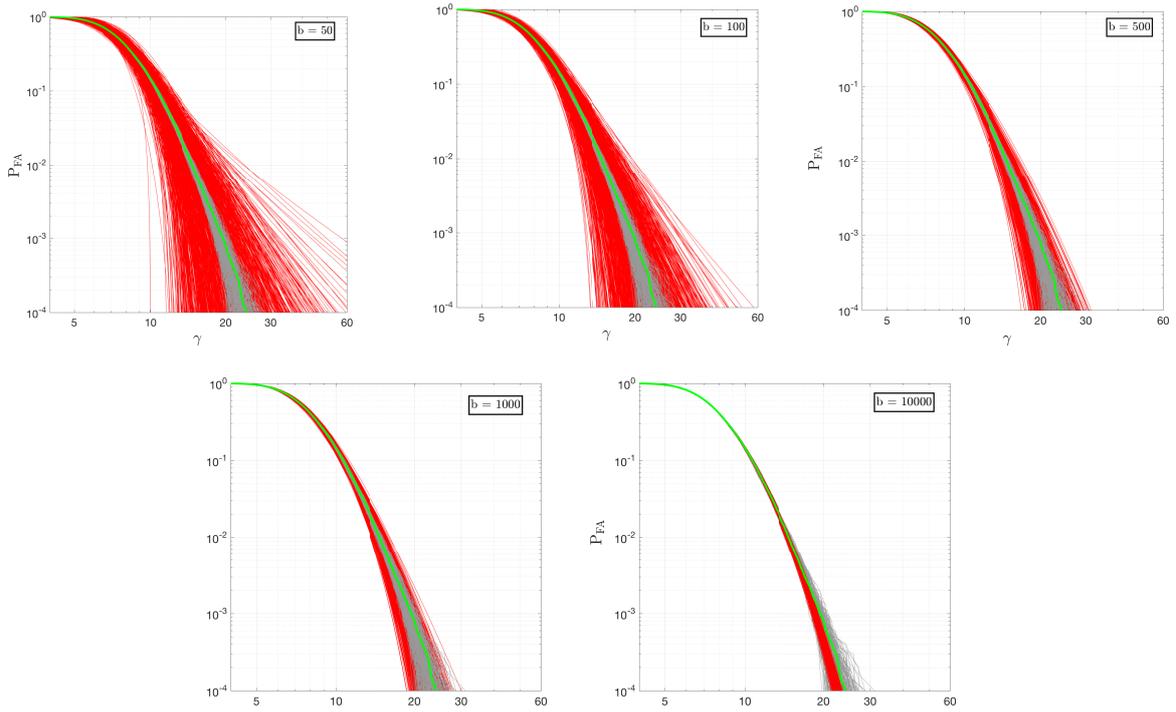


Figure 4.15 – \widehat{P}_{FA} vs P_{FA} for the proposed B_0 (grey) and B^* (red). Each panel indicates a different value of b used in B^* .

Fig. 4.16 shows $\widehat{D}_{\widehat{P}_{FA}}(B^*)$ resp. from left to right for a true $P_{FA} = 90\%$, 50% , 10% and 1% corresponding to the value of γ_0 indicated in the top row. One can see, in lexicographic order, the result when using $b = 50, 100, 500$ and 1000 MC simulations. In the three last rows, a zoom has been done for visibility. We see that the empirical distributions of the P_{FA} estimates obtained by B^* are still concentrated around the true values. As for B_0 , this means that we can use B^* in practice to evaluate the P_{FA} associated to any γ for test T_M , with only 1 set \mathcal{T}_L available. In the fourth row (corresponding to $b = 1000$), the distribution obtained with B_0 (using also

$b = 1000$) is shown in pink for comparison. One observes a narrower distribution of the FA estimates obtained using B_0 , which is due to the fact that the estimated GEV parameters in B^* depends on the number of used maxima. When b is larger, the dispersion of the FA estimates obtained with both procedures is comparable (cf. last panel of Fig. 4.15).

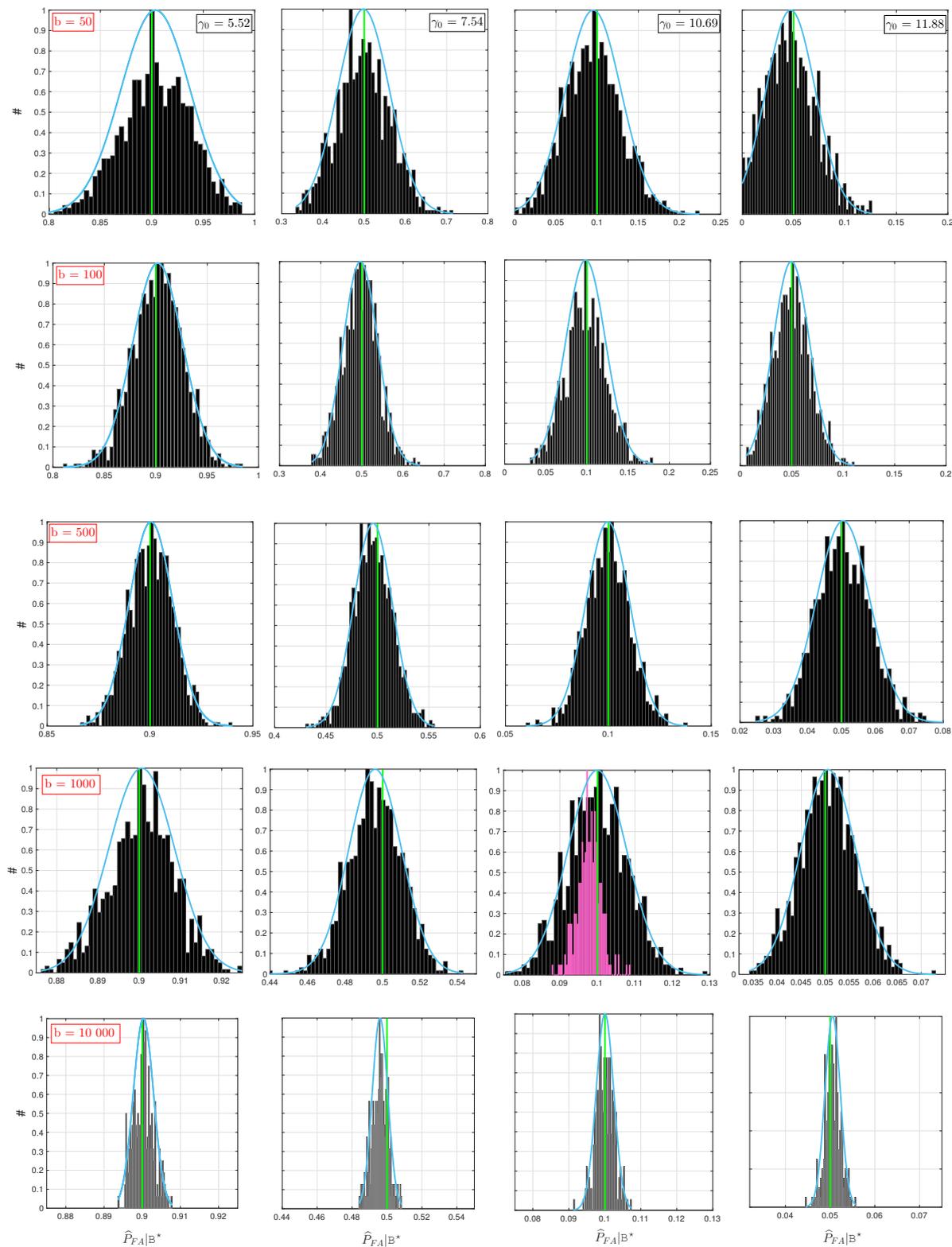


Figure 4.16 – $\widehat{D}_{\widehat{P}_{FA}}$ obtained by B^* for a true $P_{FA} = 90\%$, 50% , 10% and 1% corresponding to the value of γ_0 indicated in the top row (resp. left to right) and for $b = 50, 100, 500, 1000$ and $10\ 000$ MC simulations used (resp. from top to bottom).

Bias, variance and computational gain in time

The left panel of Fig. 4.17 compares, as a function of b , the empirical means (solid lines, right ordinate axis) and standard deviations (dashed lines, left ordinate axis) of the distributions of the FA estimates by three procedures when the test T_M in (4.21) is ran with a threshold γ corresponding to a true FA rate of 0.1. The three procedures are:

- The true distribution of the FA estimates obtained following the scheme of Fig. 4.4, assuming $L \gg 1$ sets \mathcal{T}_L are available: $\mathcal{D}_{\hat{P}_{FA}}$ (green),
- The estimated distribution of the FA estimates using procedure B_0 : $\hat{\mathcal{D}}_{\hat{P}_{FA}}(B_0)$ (black),
- The estimated distribution of the FA estimates using procedure B^* : $\hat{\mathcal{D}}_{\hat{P}_{FA}}(B^*)$ (red),

All procedures lead to a decreasing dispersion and bias as b increases, with similar estimation performance for $\mathcal{D}_{\hat{P}_{FA}}$ and $\hat{\mathcal{D}}_{\hat{P}_{FA}}$ (respectively in black and green), although for the latter only one genuine set \mathcal{T}_L is available. Moreover, these results show that the lowest dispersion and bias for $P_{FA} = 0.1$ is obtained by the third method (B^*), because the GEV model is indeed appropriate and has far less degrees of freedom than the non parametric estimate $\hat{\Phi}_M$ in (4.23) used in the two other bootstraps.

The right panel compares possible compromises dispersion *vs* computation time achievable by B_0 and B^* . For example, for $b = 1000$ we need around 14 hours of calculation (on a 4 sequential job in a 3.2 GHz processor) for both procedures and we obtain similar dispersion in terms of FA (i.e. 2.2×10^{-2} for B_0 and 1.8×10^{-2} for B^*). For $b = 10^4$ however, the computational time is much larger, about 140 hours, and it does not significantly decrease the FA dispersion (namely 10^{-2} for B_0 and 0.58×10^{-2} for B^*). To save a large amount of computational time, it appears more interesting to apply B^* with $b = 1000$ than using B_0 with $b = 10^4$ MC simulations (e.g. here, we observe a gain of $\frac{14 \times 100}{140} = 10$ at a cost of a double dispersion, which is still inferior to 10^{-2}). In this example, a compromise on the b value appears to save time while keeping a good accuracy in the P_{FA} estimates. The value of $b = 10^3$ seems to be a fair compromise.

To conclude this section, we see that B^* allows for better compromises when comparing the procedures at different b , and with a $\approx 20 - 40\%$ lower dispersion than B_0 at the same computational cost. The use of the GEV model is thus efficient to save time for the evaluation of the P_{FA} while keeping the same reliability.

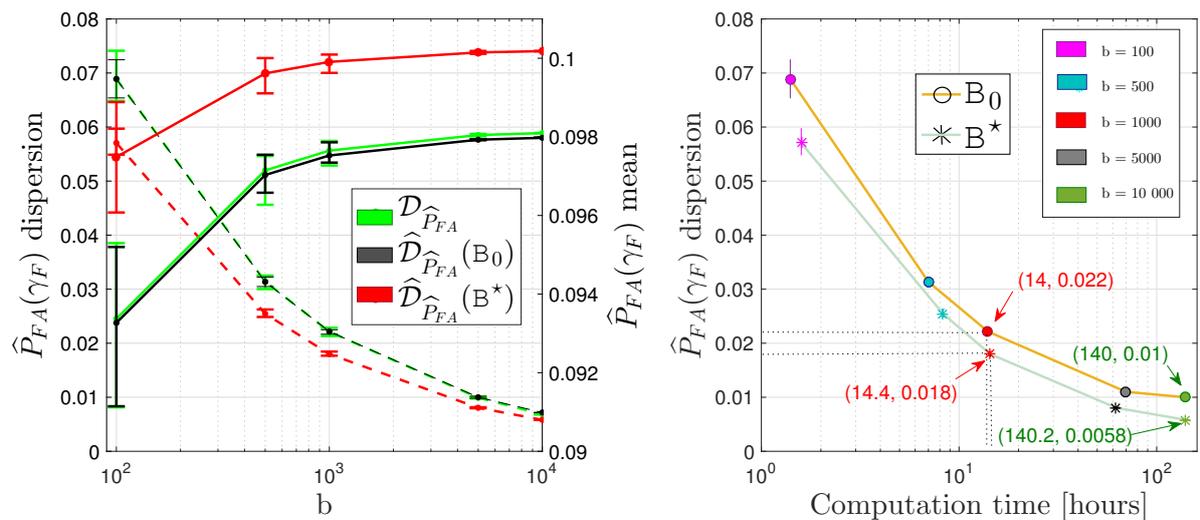


Figure 4.17 – Comparison between the B_0 and B^* procedures. Left: Comparison in terms of \hat{P}_{FA} dispersion and mean as a function of the number of used MC simulations b . Right: Dispersion of \hat{P}_{FA} as a function of the computation time.

4.7 Conclusions

In this chapter, we were interested in the control of the FA rates in case of colored noise in irregularly sampled time series. In the first sections, we have seen that various periodogram variants have been introduced in the literature to deal with the irregularity of the data sampling, which introduces spurious peaks in the periodograms.

When the noise is white Gaussian, the distribution of test statistics based on periodogram variants ordinates like the LS periodogram is difficult to evaluate as these ordinates are dependent. Several methods propose to estimate a number of “effective independent” periodogram ordinates to derive the test statistics distribution. Although such empirical methods to estimate the FA rates are largely used in the literature, their reliability is strongly debated even for the simplest case of WGN. To control the FA, we can instead turn to bootstrap analyses and to extreme value theory to model the test statistics distribution.

In case of colored noise, several methods consist in modelling the noise correlations through parametric models. They produce FA estimates that are always model dependent.

In Sec. 4.5, we have investigated the possibility of using noise training data sets to improve the control of the false alarm rate associated to detection test in the case of irregular sampling. The proposed method is based on standardized periodograms (which have been considered in Chap. 3 in the case of regularly sampled observations). By developing an adapted bootstrap procedure in the Fourier domain, it appears that one can reliably bound the false alarm probability corresponding to any threshold for test $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ (4.21) in the general case of irregularly sampled observations by taking advantage from a training data set of the noise. However, as this procedure is based on MC simulations, it is computationally heavy. We showed that exploiting a GEV distribution allows to save time while keeping the same behavior for the resulting FA estimates (cf. Sec. 4.5.2 and Sec. 4.6.2). Of course, this general procedure can be useful for the detection of extrasolar planets in RV data: in this case, the time sampling is often uneven, the exoplanetary signatures is hidden in the colored noise coming from the stellar surface convection, and this noise can be accurately simulated using astrophysical codes.

Conclusions and perspectives

Summary and main contributions

This thesis work was first motivated by its application to exoplanet detection by the radial velocity technique. In this spirit, we have dedicated the first part of the Introduction (Chap. 1) to the astrophysical problem of this thesis: controlling the significance level of planets detection tests when planets' signatures are hidden in the colored stellar noise. In parallel, 3D hydrodynamic simulations of the stochastic convective noise are in rapid development since the last decades, and they are now able to produce reliable time series of this “noise” source. In the aim of using such noise training data set in the detection process, we have then dedicated this thesis to the development of statistical methods taking advantages of such simulations of the stellar noise.

Before getting to the heart of this research work, we tried to provide in a first chapter a detailed introduction to the astrophysical context of extrasolar planet detection and stellar noise to readers more familiar with detection theory. In a second chapter, we tried to provide a detailed introduction of detection theory to readers more familiar with the astrophysical context. In this second chapter, we have reviewed some bases of hypothesis testing methodology, the periodogram properties as a PSD estimate, and different detection tests. Through the investigation of the RV planetary signature in the periodogram, we have seen that the planetary signals are sparse for the majority of the planet (Keplerian) parameters. This allowed to set up a detection strategy based on a simplified sinusoid-based model and on a well understood and efficient tool for sinusoid detection, the periodogram. We have studied the statistical (asymptotic) properties of different test statistics applied to the periodogram for different cases of noises (white and colored) under the null, and different signals (single or multiple) under the alternative. We have seen that the false alarm and detection probabilities can not be known when the noise is colored with unknown statistics for the considered tests from the literature.

As a very general problem, we were then interested in Chap. 3 to investigate whether noise training data sets could improve the detection process and in particular allow the control of the FA rates of some of the considered tests statistics. In this spirit, we have designed a standardized periodogram, which consists in normalizing the periodogram of the data by an averaged periodogram coming from available noise training data set. The statistical analysis of the standardized periodogram shows that $\tilde{\mathbf{P}}|\overline{\mathbf{P}}_L$ is asymptotically independent of the noise PSD under the null. This property allowed us to characterize the FA and detection probabilities of different detection tests, which are CFAR thanks to the standardization. In particular, we paid special attention to:

- The test based on the maximum periodogram value ($T_M(\tilde{\mathbf{P}}|\overline{\mathbf{P}}_L)$). This test is the most powerful when the signal is extremely sparse under the alternative. Analytical expressions of FA and detection probabilities have been derived and are given in (3.18) and (3.19).
- The multisine test based on the N_C^{th} maximum standardized periodogram value ($T_C(\tilde{\mathbf{P}}|\overline{\mathbf{P}}_L)$), where N_C is the *a priori* number of harmonics. For this test, we also derived

analytical expressions for both probabilities, which are given in (3.38) and (3.41). We have seen that this test can outperform $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ when N_S is large. We have also seen that the performances of this test varies only slightly with the setting parameter N_C , which makes it very interesting when we can expect a large (w.r.t. 1) number of periodogram components contaminated by the periodic signal.

- The adaptive tests based on the standardized periodogram P -values ($HC(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$ and $BJ(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$). For these tests, which are also CFAR thanks to the proposed standardization, P -values have to be computed through numerical methods. We have seen that, in case of non extremely sparse signal, these tests have excellent performances w.r.t. T_M and even T_C . Such adaptive tests are particularly interesting as they do not need any *a priori* on the number of sinusoidal components. In particular, BJ allows to efficiently detect the correct number of components at their true location.

Although our results are asymptotic, numerous simulations showed that they accurately reflect the tests performances in terms of false alarm and detection rates. These results are also quite general in the sense that they can be applied to any sinusoid detection problem for which (even a few) training data sets are available. One has observed that the performances of any test applied to $\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L$ increase with the number of used noise time series. The statistical study of these tests statistics opens the door to various studies. Indeed, we can:

- Derive reliable detection thresholds in case of colored noise,
- Compare the different tests' performances (e.g. through ROC curves) for various types of alternatives,
- Design efficient detection strategies for various detection tests to found the best observational parameters (N , Δt) to be able to detect a given signal with high probability.

Considering then the case of time series irregularly sampled in Chap. 4, we have seen that the periodogram ordinates are strongly dependent, making analytical expressions of the FA rates out of reach. This is also true for the different periodograms used in practice as the Lomb-Scargle periodogram or its generalizations. We proposed an adapted bootstrap procedure to control the FA rates. This procedure is also based on the noise training data sets. Applied to a synthetic colored noise, this bootstrap method is efficient to evaluate the true false alarm rate at which the detection test (e.g. T_M) is conducted for all types of data sampling. It is quite adaptable, as different periodograms can be used. Moreover, while the procedure has been illustrated on test $T_M(\tilde{\mathbf{P}}|\bar{\mathbf{P}}_L)$, the implementation of other test statistics is straightforward. Considering the case where we dispose of only a few noise training data sets (e.g. due to a large computational time to generate them), the procedure involves a parametric estimation for which the noise model can be well chosen (e.g. by checking our model choice through diagnostic plots). In practice, AR colored noise could be a good choice as it is a quite general colored noise model for which different AR estimation criteria exist, but any parametric noise model can be chosen. Moreover, as the estimation is done on the simulated noise training data set, the estimation is not influenced by missing data during the estimation process.

As a lot of colored noise realizations have to be performed during the proposed bootstrap procedure, we have seen that using GEV model allows to save an important amount of computational time. Making this approximation is at the price of a slightly larger FA estimates interval.

Using independent noise training data sets in the detection process using the proposed standardized periodogram presents several advantages:

- First, comparing with classical methods ignoring the noise correlations, it allows to control the FA rates independently of the colored noise PSD.

- It does not need to make an *a priori* decision about the periodogram components contaminated by the noise, as in the CLEANing methods.
- The use of the noise training data set allows to control the FA and detection probabilities by simple analytical expressions in the case of a regular sampling and by bounded estimates in the case of irregularly sampled observations.
- The noise simulations constitute then a useful tool to design detection strategies for a given signal to choose the best observational parameters and detection test.

Application to exoplanet detection

The statistical methods developed in Chap. 3 and 4 are based on the knowledge of the colored noise. In these chapters, we used autoregressive processes for the noise in order to easily generate time series and easily verify the developed theoretical expressions. The following step is to apply 3D HD simulations to standardize the periodogram of RV data. Therefore, we aim to compare the realism of such 3D HD simulations with real observations of the stellar convection. To do so, the Sun is an ideal target since very long (quasi-regularly sampled) time series of the solar RV are available.

We consider the RV data taken by *Global Oscillations at Low Frequency* (GOLF) resonant scattering spectrophotometer on board the SoHo satellite⁶, which observes the Sun for two decades (i.e. almost 2 solar cycles). This instrument observes the neutral sodium doublet lines \mathcal{D}_1 at $\lambda = 589.6$ nm and \mathcal{D}_2 at $\lambda = 589.0$ nm. The RV extraction method is different from the cross-correlation with ground-based spectrographs such as HARPS or HARPS-N (described in Sec. 1.2.2). Spectrophotometers like GOLF extract the velocities by measuring the intensity contrast in the observed absorption lines (for a detailed description of this extraction method, one can report to [Boumier and Dame, 1993, Gabriel et al., 1995, Garcia et al., 2005]). We use RV time series calibrated by Garcia et al. and filtered at 3 days to remove the long term magnetic activity contributions [Garcia et al., 2005]. It should contain therefore only the contribution of the solar convection and oscillations modes. We selected a time series of $T = 54$ days without any gaps in the years 1996-1997 to be close to the solar cycle minimum (to minimize any potential contribution of magnetic activity). However, one can note that Garcia et al. compared the solar PSD at different epochs of the solar cycle and did not find significant variations with the magnetic cycle (for similar conclusions, see also [Lefebvre et al., 2008]).

To simulate the surface convection and stratification of the upper layers of the Sun, we use the state-of-the-art radiative hydrodynamical: the STAGGER CODE⁷, cf. [Nordlund and Galsgaard, 1995]. In a 3D local-box model of the solar photosphere⁸, the code solves the full set of conservative hydrodynamical equations coupled to an accurate treatment of the radiative transfer. The code is based on a sixth order explicit finite difference scheme. The equations are solved on a staggered mesh where the thermodynamical variables are cell centred, while the flux are shifted to the cell edge. The domain of simulation contains the entropy minimum located at the surface and is extended deep enough to have a flat entropy profile at the bottom (adiabatic regime). The code uses periodic boundary conditions horizontally and open boundaries vertically. At the bottom of the simulation, the inflows have constant entropy and pressure. The outflows are not constrained and are free to pass through the boundary. We used a realistic equation-of-state that accounts for ionization, recombination, and dissociation, and continuous plus line opacities. Radiative transfer is crucial since it drives convection through entropy losses at the surface [Stein and Nordlund, 1998]. The wavelength dependence of the radiative transfer is taken into account using a binning scheme, in which the

⁶http://irfu.cea.fr/Sap/Phoccea/Vie_des_labos/Ast/ast_technique.php?id_ast=1130

⁷<http://www.astro.ku.dk/~kg/Papers/MHDcode.ps.gz>

⁸The box size is: 8000 km² and +500 and -3400 kms above and below the surface at $\tau = 1$.

monochromatic lines are collected into 12 bins. The numerical resolution used for the present simulation is 120^3 . The choice of this modest resolution is a compromise between sufficient fine grid to catch enough of the inhomogeneities and sufficiently small to minimize the computing and storing costs of very long run simulation. Several runs have been performed with different magnetic fields injected in the simulation. In order to compare our RV to GOLF data, we compute synthetic line profiles of the Sodium doublet from 3D snapshots every minutes and for a sequence of 54 days. The synthetic disk-integrated intensity line profiles of the considered lines have been computed at 6 different angles from disk center to the limb (using a Radau integration). An example of such synthetic spectrum is shown in Fig. 4.5.

The RV is extracted from the line profiles by the intensity contrast in the blue (I_B) and red (I_R) wings, related to the RV by:

$$V_R(t) = \kappa \frac{I_B - I_R}{I_B + I_R}, \quad (4.25)$$

with κ a proportionality term related to the slope at the wavelength where the RV is extracted (cf. p.328 of [Unno et al., 1989] and to Sulis. et al., in prep).

The left panel of Fig. 4.5 shows a synthetic spectrum of the Na lines (called \mathcal{D}_1 and \mathcal{D}_2) at a given time. The right panel illustrates the velocity extraction method on one line. The reference spectrum is shown in black and the spectrum measured at a given time t , $I(t, \lambda)$, is shown by the dot green line. In this approach, the wavelength shift ($\Delta\lambda$) is not the measured quantity but rather the change of intensity at fixed wavelength (λ_0^B, λ_0^R). For small shifts, $\Delta\lambda$ and ΔI are related by Taylor development, which leads to expression (4.25).

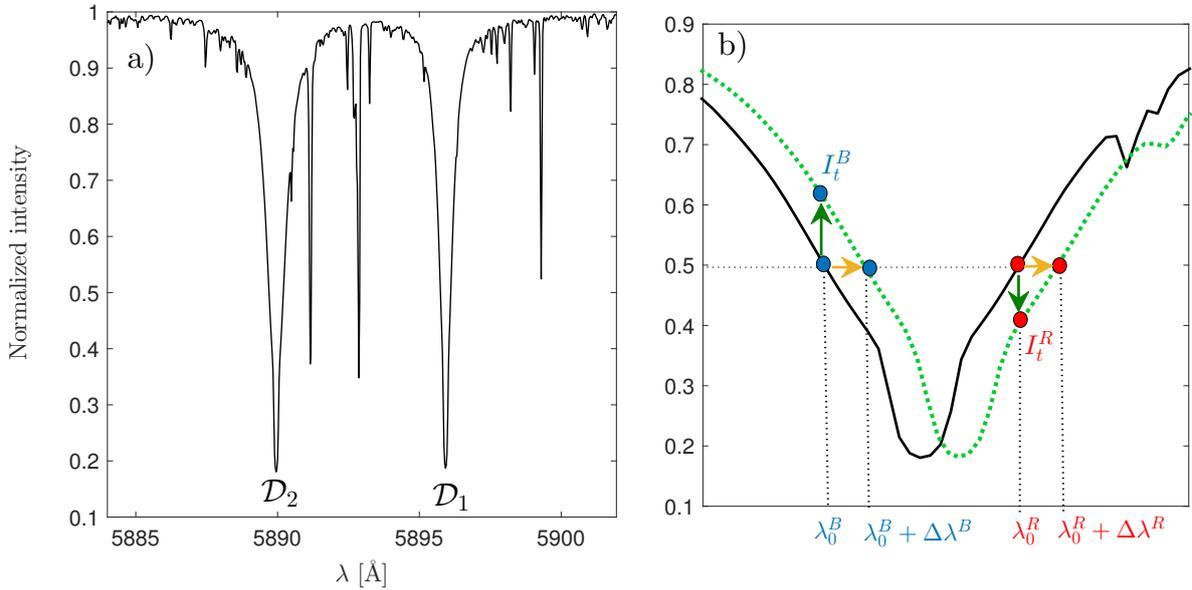


Figure 4.5 – Simulated spectrum of the solar sodium doublet (left) and blueprint the velocity extraction method on one line (right). The intensity changes are measured by GOLF at two fixed wavelength (λ_0^B and λ_0^R).

The comparison of the simulated RV extracted from time series of 3D spectra (red) with GOLF observations (black) is shown in Fig. 4.6. The left panel compares the RV and the right panel the smoothed (of 10 bins) periodograms. We see a good match between the observations and the simulations until $\approx 10^{-5}$ Hz corresponding to a period of the day. At the lower frequency, the observations are filtered and the comparison is not possible (shade region). Both the simulations and the observations show that the granulation contribution cannot be treated as a WGN (flat power over the whole frequency range). The convective noise is frequency dependent over all the frequencies. It increases towards the lower frequencies. The rms of radial velocities of the observations is 1.5 m.s^{-1} and 0.95 m.s^{-1} for the simulated velocities (which is closed to the value obtained by [Meunier et al., 2015] for the RV rms of the granulation, i.e. around 0.8 m.s^{-1}). Another interesting finding is that the PSD is dominated by the horizontal component of the velocity, as noticed also in [Meunier et al., 2015]. Finally, we note that a transverse magnetic field of 150 Gauss has been used in the simulations but we will explore in details different strength fields for a better matching [Sulis et al., in preparation]. Indeed, the magnetic field might affect the RV amplitudes and explain the slight disagreements shown in Fig. 4.6 (left).

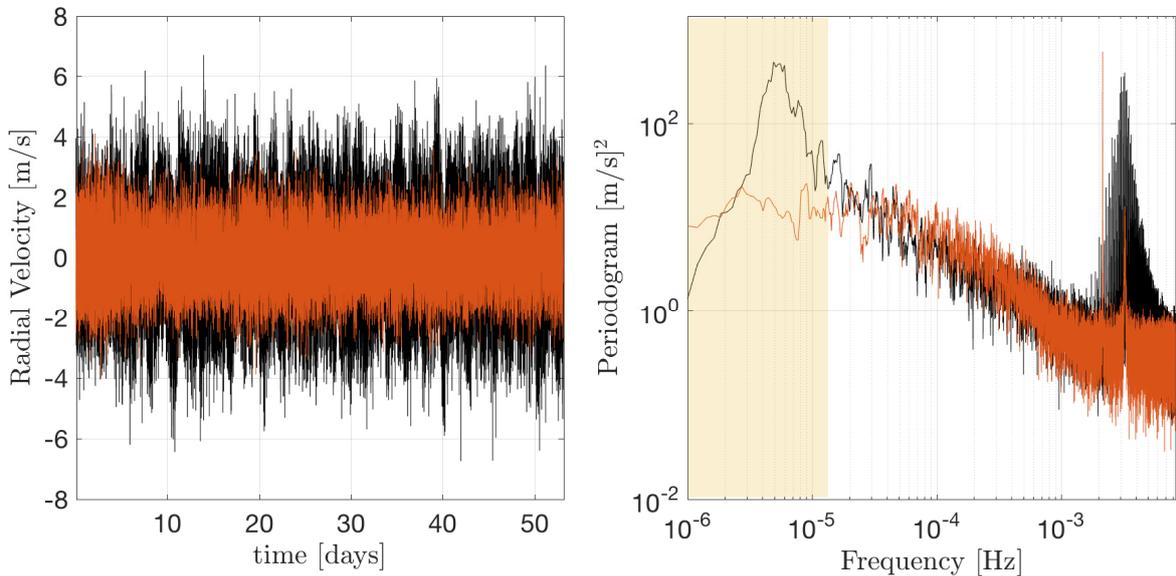


Figure 4.6 – Left: Comparison of solar data (black) with the RV extracted from disk integration of synthetic sodium spectral lines \mathcal{D}_1 and \mathcal{D}_2 from hydrodynamical simulations of granulation (red). Right: Smoothed periodograms of GOLF data (black) and simulated data (red). The shade region illustrates the frequency range where the data have been filtered [Garcia et al., 2005]. To the simulated RV time series, we have added a WGN of variance 2 to mimic the instrumental noise at highest frequencies. This instrumental noise does not affect the lower frequency part of the periodogram.

This preliminary comparison will be published in A&A [Sulis et al., in prep.]. In this paper, we aim to focus on three main objectives:

- (i) Show that 3D HD simulations are sufficiently realistic to reproduce the convective noise in a Sun-like star,
- (ii) Show the advantages of using the *ab initio* HD simulations to control the detection claims by injecting false planetary signals in the solar data.
- (iii) Study the planet detectability in presence of the convective noise source only. Indeed, applying the periodogram standardization given in (3.2) using HD simulations will allow to design detection strategies as in Sec. 3.5.5. Considering test T_M in (3.11) with P_{FA}

given in (3.18) and P_{DET} in (3.19), the application is straightforward. One can evaluate several planetary (e.g. planet mass and period) and observational (e.g. time sampling) parameters to predict the performances of the test as a function of the time of observation.

Moreover, using the RV resulting from the 3D HD simulations, we aim to

1. Study the influence of the magnetic field on the convective part of the solar DSP to find if the simulations are in agreement with the results of [Garcia et al., 2005] (who found no significant variations with the magnetic stellar cycle).
2. Investigate the influence of the fundamental stellar parameters on the convective PSD. Indeed, in practice, the measures of the stellar fundamental parameters are always given with incertitudes and we have to know the impact on the resulting PSD shape.
3. Compare the detection significances derived by classical methods with our method for some small planetary signals at short periods.
4. Design detectability studies for different kind of stars, samplings and target planetary signatures to evaluate the best observational strategies to deal with the stellar convective noise source.

Further interesting perspectives for the long term

This statistical study coupled with the reliability of the HD simulations to describe the convective noise open also the door to further astronomical studies.

Adaptive tests. First, it should be interesting to investigate the performances of adaptive tests in the case of irregular sampling. Indeed, as discussed quickly in Chap. 4 (cf. Sec. 4.1), the detection tests performances can be better for irregular sampling than for regular ones. The relative behavior of these tests is unknown for the moment when the sampling is irregular and their performances in case of non extremely sparse signal (e.g. high eccentric planet orbit) should be investigated.

Long computation time of the simulations. As the recent HD simulations are heavy in computation time, one can imagine to produce “reference tables” of simulated data with a very fine sampling (to be able to cope with the case of irregular sampling) for different kind of stars.

Contribution of other noise sources. The contribution of the other noise sources that cannot be taken into account through the simulated noise training data set should also be investigated (e.g. instrumental noise or other noise sources as the magnetic activity). Different approaches have to be considered depending on the other noise sources characteristics. Are they stochastic or determinist ? Is it possible to produce independent noise training data sets of such noises (e.g. instrumental noises, spot contribution correlated to the frequency of the stellar surface rotation) ? Moreover, if we have access to external indicators (e.g. chromospheric indicators) in order to remove their contributions at specific frequencies (through a CLEAN procedure), these specific frequencies should not be considered to derived the FA rate and the proposed periodogram standardization should perhaps be modified.

Planetary transits. Then, the developed methodology based on noise training data sets should also be investigated to the planet detection by the transit methods. In this case, the periodogram is no more appropriate as the signature is spread in the temporal domain (this signature is indeed periodic but we observe very few periods in general), and with a

shape which is very different from a sinusoidal form. In this case, classical methods consist in taking advantage from the shape of the signature of interest and use specific *matched filters* to represent the signal. The main detection algorithm consists in comparing the light curves with a transit model for different parameters (epoch of the transit beginning, period and duration). For example the *Box-Least-Square* algorithm [Kovacs et al., 2002] consists in finding the best fit using mean square errors with a rectangular function for the transit. However, the stellar noise is not taken into account and the FA cannot be reliably derived when the noise is colored, as for the RV detection problem. Some matched filter methods consist in *whitening* the noise with adapted filters but are based on the observed data and it is difficult to control the statistical properties of the “whitened” estimate. In this context, the proposed methodology could take advantage from independent noise training data set coming from HD simulations coupled with a matched filter approach, and with detection tests dedicated to these deviations of weak amplitudes in massive time series. In this case also, we can expect that the use of simulated photometric time series coming from HD simulations of the granulation phenomenon could be very useful to control the tests’ significance level. We note that the use of HD simulations has already been investigated on photometric data in [Chiavassa et al., 2015, Chiavassa et al., 2017] to highlight that a full characterization of the granulation noise should be considered for detecting (and characterizing) the planetary transits.

Appendices

Appendix A

Theoretical distributions

This appendix lists the main theoretical distributions discussed in this work. One can report to [Abramowitz et al., 1972] for further characteristics of these distributions.

Gaussian distribution

The probability density function of a normal r.v. with mean μ and variance σ^2 is:

$$\varphi_G(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad (\text{A.1})$$

for $x \in \mathbb{R}$, $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.

The corresponding cumulative distribution function is

$$\Phi_G(x; \mu, \sigma^2) = \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{x - \mu}{\sigma\sqrt{2}} \right) \right], \quad (\text{A.2})$$

with $\operatorname{erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$ the error function.

	Mode $m[x]$	Mean $\mathbb{E}[x]$	Variance $\operatorname{var}[x]$
Exists for	$\mu \in \mathbb{R}$	$\mu \in \mathbb{R}$	$\sigma^2 > 0$
Expression	μ	μ	σ^2

Table A.1 – Some properties of the Gaussian distribution.

χ^2 distribution

The probability density function of a chi-square r.v. with d degrees of freedom is:

$$\varphi_{\chi^2}(x; d) = \frac{x^{d/2-1} e^{-x/2}}{2^{d/2} \Gamma(\frac{d}{2})}, \quad (\text{A.3})$$

for $x > 0$ and is null otherwise. The Gamma function is defined as $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$.

The corresponding cumulative distribution function is

$$\Phi_{\chi^2}(x; d) = \Gamma_R \left(\frac{d}{2}, \frac{x}{2} \right) \quad (\text{A.4})$$

with $\Gamma_R(d, x)$ the regularized Gamma function.

	Mode $m[x]$	Mean $\mathbb{E}[x]$	Variance $\text{var}[x]$
Exists for	$d \geq 2$	$d > 0$	$d > 0$
Expression	$[0, d - 2]_+$	d	$2d$

Table A.2 – Some properties of the χ^2 law.

Non central χ^2 distribution

The probability density function of a non central chi-square r.v. with d degrees of freedom and non centrality parameter λ is:

$$\varphi_{\chi^2}(x; d; \lambda) = \frac{1}{2} e^{-(x+\lambda)/2} \left(\frac{x}{\lambda}\right)^{d/4-1/2} I_{d/2-1}^B(\sqrt{\lambda x}) \quad (\text{A.5})$$

with I_a^B the modified Bessel function of the first kind defined as $I_a^B(b) = (b/2)^a \sum_{j=0}^{\infty} \frac{(b^2/4)^j}{j! \Gamma(a + j + a)}$.

The corresponding cumulative distribution function is

$$\Phi_{\chi^2}(x; d, \lambda) = e^{-\lambda/2} \sum_{j=0}^{\infty} \frac{(\lambda/2)^j}{j!} \Gamma_R(x, d + 2j). \quad (\text{A.6})$$

	Mode $m[x]$	Mean $\mathbb{E}[x]$	Variance $\text{var}[x]$
Exists for	$k, \lambda > 0$	$k, \lambda > 0$	$k, \lambda > 0$
Expression	(*)	$d + \lambda$	$2(d + 2\lambda)$

Table A.3 – Some properties of the non central χ^2 law. (*) The mode has to be evaluated numerically.

F distribution

The probability density function of a F r.v. with d_1, d_2 degrees of freedom is:

$$\varphi_F = \frac{\sqrt{\frac{(d_1 x)^{d_1} (d_2)^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}, \quad (\text{A.7})$$

with the Beta function defined as $B(x, y) = \int_0^1 x^{t-1} t^{y-1} (1-t)^{y-1} dt$.

The corresponding cumulative distribution function is

$$\Phi_F(x, d_1, d_2) = I_{\frac{d_1 x}{d_1 x + d_2}}\left(\frac{d_1}{2}, \frac{d_2}{2}\right), \quad (\text{A.8})$$

with $I_x(a, b)$ the regularized incomplete beta function, defined by the ratio of the incomplete beta function $B(x; a, b)$ and the complete beta function $B(a, b)$:

$$I_x(a, b) = \frac{B(x; a, b)}{B(a, b)} = \frac{1}{B(a, b)} \int_0^x t^{a-1} (1-t)^{b-1} dt.$$

If $a, b \in \mathbb{N}$, we can use integration by parts to work out the integral. Then, we can develop the expression in terms of binomial expansion (given by Eq 6.6.4, p263 of [Abramowitz et al., 1972]) as:

$$I_x(a, b) = \sum_{j=a}^{a+b-1} \frac{(a+b-1)!}{j!(a+b-1-j)!} x^j (1-x)^{a+b-1-j}.$$

	Mode $m[x]$	Mean $\mathbb{E}[x]$	Variance $\text{var}[x]$
Exists for	$d_1 > 2$	$d_2 > 2$	$d_2 > 4$
Expression	$\frac{d_1-2}{d_1} \frac{d_2}{d_2+2}$	$\frac{d_2}{d_2-2}$	$\frac{2d_1^2(d_2+d_1-2)}{d_2(d_1-2)^2(d_1-4)}$

Table A.4 – Some properties of the central F law.

Non central F distribution

The probability density function of a non central F r.v. with d_1, d_2 degrees of freedom and non centrality parameter λ is:

$$\varphi_{F_\lambda}(x, d_1, d_2) = \sum_{k=0}^{\infty} \frac{e^{-\frac{\lambda}{2}} (\lambda/2)^k}{B(\frac{d_2}{2}, \frac{d_1}{2} + k) k!} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2} + k} \left(\frac{d_2}{d_2 + d_1 x}\right)^{\frac{d_1 + d_2}{2} + k} x^{\frac{d_1}{2} - 1 + k} \quad (\text{A.9})$$

for $x \geq 0$ and zero otherwise and $d_1, d_2, \lambda > 0$.

The corresponding cumulative distribution function is

$$\Phi_{F_\lambda}(x, d_1, d_2) = \sum_{k=0}^{\infty} \left(\frac{(\frac{1}{2}\lambda)^k}{k!} e^{-\frac{\lambda}{2}} \right) I_{\frac{d_1 x}{d_1 x + d_2}} \left(\frac{d_1}{2} + k, \frac{d_2}{2} \right). \quad (\text{A.10})$$

When $\lambda = 0$, the non central F distribution becomes a central F distribution.

	Mode $m[x]$	Mean $\mathbb{E}[x]$	Variance $\text{Var}[x]$
Exists for	$d_1, d_2, \lambda > 0$	$d_1, \lambda > 0, d_2 > 2$	$d_1, \lambda > 0, d_2 > 4$
Expression	(*)	$\frac{d_2(d_1+\lambda)}{d_1(d_2-2)}$	$2 \frac{(d_1+\lambda)^2 + (d_1+2\lambda)(d_2-2)}{(d_2-2)^2(d_2-4)}$

Table A.5 – Some properties of the non-central F law. (*) The mode has to be evaluated numerically.

Appendix B

Periodogram distribution and non centrality parameter

Derivation of expressions (2.39) and (3.6)

We prove here that for model (3.1) the periodogram is asymptotically distributed as in (2.39) with non centrality parameters as in (3.6). The proof is adapted from Theorem. 2 extracted from [Li, 2014], which considers the complex case. We first prove (2.39) and then turn to (3.6).

The time series of model (3.1) can also be written as

$$\begin{aligned} X(j) &= \sum_{q=1}^{N_s} \alpha_q \sin(2\pi f_q j + \varphi_q) + E(j) \\ &= R(j) + E(j), \end{aligned}$$

with $R(j) := \sum_{q=1}^{N_s} \alpha_q \sin(2\pi f_q j + \varphi_q)$ a deterministic part, which using Euler formulae can be written as

$$R(j) = \sum_{q=1}^{N_s} \frac{\alpha_q}{2} e^{i(\varphi_q - \frac{\pi}{2})} e^{2\pi i f_q j} - \frac{\alpha_q}{2} e^{-i(\varphi_q + \frac{\pi}{2})} e^{-2\pi i f_q j}.$$

By introducing:

$$\begin{aligned} \mathbf{g}(\nu) &:= [e^{i2\pi\nu}, \dots, e^{i2\pi N\nu}]^\top, \\ \mathbf{g}^+(f_q) &:= e^{i(\varphi_q - \frac{\pi}{2})} \mathbf{g}(f_q), \\ \mathbf{g}^-(f_q) &:= e^{-i(\varphi_q + \frac{\pi}{2})} \mathbf{g}(f_q), \end{aligned}$$

the time series writes in vector form:

$$\mathbf{X} = \sum_{q=1}^{N_s} \frac{\alpha_q}{2} \left(\mathbf{g}^+(f_q) - \mathbf{g}^-(f_q) \right) + \mathbf{E} = \mathbf{R} + \mathbf{E} \quad (\text{B.1})$$

and its discrete Fourier transform (DFT) y_k at frequency ν_k writes

$$y_k = \frac{1}{N} \mathbf{g}^H(\nu_k) \mathbf{X} = \frac{1}{N} \mathbf{g}^H(\nu_k) \mathbf{R} + \frac{1}{N} \mathbf{g}^H(\nu_k) \mathbf{E}.$$

The DFT is composed of a deterministic part, μ_k , and a stochastic part, ϵ_k , defined as

$$\mu_k := \frac{1}{N} \mathbf{g}^H(\nu_k) \mathbf{R} \quad \text{and} \quad \epsilon_k := \frac{1}{N} \mathbf{g}^H(\nu_k) \mathbf{E}. \quad (\text{B.2})$$

Because E is a zero mean Gaussian process, the random variable $y_k = \mu_k + \epsilon_k$ is Gaussian with mean μ_k and variance σ_k^2 . This is a complex variable for all Fourier frequencies except for ν_0 and $\nu_{\frac{N}{2}}$ because $\mathbf{g}(\nu_0)$ and $\mathbf{g}(\nu_{\frac{N}{2}})$ are real.

The distribution of the periodogram requires to investigate the variance of ϵ_k which, with (B.2), writes:

$$\begin{aligned}
N\text{Var}[\epsilon_k] &= N^{-1}\mathbb{E}\left(\mathbf{g}^H(\nu_k)\mathbf{E}\mathbf{E}^\top\mathbf{g}(\nu_k)\right) \\
&= N^{-1}\sum_{t,s=1}^N r_E(t-s)e^{-i2\pi\nu_k(t-s)} \\
&= N^{-1}\sum_{|u|<N} (N-|u|)r_E(u)e^{-i2\pi\nu_k u} \\
&= S_E(\nu_k) - \sum_{|u|<N} N^{-1}|u|r_E(u)e^{-i2\pi\nu_k u} - \sum_{|u|\geq N} r_E(u)e^{-i2\pi\nu_k u} \\
&= S_E(\nu_k) + \mathcal{O}(r_N),
\end{aligned} \tag{B.3}$$

since, using $|r_E(u)| < \infty$ and the dominated convergence theorem we have

$$r_N := \sum_u \min\left(1, \frac{|u|}{N}\right) |r_E(u)| \rightarrow 0 \quad \text{as } N \rightarrow \infty. \tag{B.4}$$

Hence, for all Fourier frequencies,

$$\sigma_k^2 := \text{Var}[\epsilon_k] = N^{-1}S_E(\nu_k) + \mathcal{O}(N^{-1}r_N). \tag{B.5}$$

By Lemma 12.2.1(b) of [Li, 2014], we obtain

$$|y_k|^2/\sigma_k^2 \sim \begin{cases} \frac{1}{2}\chi_{2,2\frac{|\mu_k|^2}{\sigma_k^2}}^2, & \forall k \in \Omega, \\ \chi_{1,\frac{|\mu_k|^2}{\sigma_k^2}}^2, & \text{for } k = 0, \frac{N}{2}. \end{cases}$$

Hence, for the periodogram this implies

$$P(\nu_k|H_1) = N|y_k|^2 = N\sigma_k^2 \frac{|y_k|^2}{\sigma_k^2} \sim \begin{cases} \frac{1}{2}\rho_k^{-1}S_E(\nu_k)\chi_{2,2\rho_k\gamma_k}^2, & \forall k \in \Omega, \\ \rho_k^{-1}S_E(\nu_k)\chi_{1,\rho_k\gamma_k}^2, & \text{for } k = 0, \frac{N}{2}, \end{cases} \tag{B.6}$$

where

$$\rho_k := \frac{S_E(\nu_k)}{N\sigma_k^2} \quad \text{and} \quad \gamma_k := N \frac{|\mu_k|^2}{S_E(\nu_k)}. \tag{B.7}$$

With (B.5), we see that

$$\rho_k = \frac{S_E(\nu_k)}{S_E(\nu_k) + \mathcal{O}(r_N)} = 1 + \mathcal{O}(r_N)$$

for all Fourier frequencies. Owing to (B.4), an approximated distribution of (B.6) can be obtained by neglecting the $\mathcal{O}(r_N)$ in (2.39). The distribution (2.39) follows by noting

$$\lambda_k := 2\gamma_k \quad \text{for } k \in \Omega \quad \text{and} \quad \lambda_k := \gamma_k \quad \text{for } k = 0, \frac{N}{2}. \tag{B.8}$$

We now turn to the computation of the non centrality parameters given in (3.6). We have from (B.1), (B.2) and (B.7)

$$\begin{aligned}\gamma_k &= \frac{N}{S_E(\nu_k)} \left| \frac{1}{N} \mathbf{g}^H(\nu_k) \mathbf{R} \right|^2 \\ &= \frac{1}{NS_E(\nu_k)} \left| \sum_{j=1}^N \sum_{q=1}^{N_s} \frac{\alpha_q}{2} \left(e^{i(\varphi_q - \frac{\pi}{2})} e^{2\pi i(f_q - \nu_k)j} - e^{-i(\varphi_q + \frac{\pi}{2})} e^{-2\pi i(f_q + \nu_k)j} \right) \right|^2.\end{aligned}\tag{B.9}$$

As a result, the non centrality parameter is:

$$\begin{aligned}\lambda_k &= 2\kappa_k \mu_k \\ &\approx \frac{2}{S_E(\nu_k)} \frac{|\text{FT}(\mathbf{R}(t_j))|^2}{N} \\ &= \frac{2}{NS_E(\nu_k)} \left| \sum_{j=1}^N R(t_j) e^{-2\pi i \nu_k t_j} \right|^2 \\ &= \frac{2}{NS_E(\nu_k)} \left| \sum_{j=1}^N \sum_{p=1}^{N_s} \frac{\alpha_q}{2} \left(e^{i(\phi_q - \frac{\pi}{2})} e^{2\pi i f_q t_j} - e^{-i(\phi_q + \frac{\pi}{2})} e^{-2\pi i f_q t_j} \right) e^{-2\pi i \nu_k t_j} \right|^2 \\ &= \frac{2}{NS_E(\nu_k)} \left| \sum_{j=1}^N \sum_{p=1}^{N_s} \frac{\alpha_q}{2} \left(e^{i(\phi_q - \frac{\pi}{2})} e^{2\pi i(f_q - \nu_k)t_j} - e^{-i(\phi_q + \frac{\pi}{2})} e^{-2\pi i(f_q + \nu_k)t_j} \right) \right|^2 \\ &= \frac{2}{NS_E(\nu_k)} \left| N \sum_{p=1}^{N_s} \frac{\alpha_q}{2} \left(D_N(f_q - \nu_k) e^{i(\phi_q - \frac{\pi}{2})} - D_N(f_q + \nu_k) e^{-i(\phi_q + \frac{\pi}{2})} \right) \right|^2 \\ &= \frac{N}{2S_E(\nu_k)} \left| \sum_{p=1}^{N_s} \alpha_q \left(D_N(f_q - \nu_k) e^{i(\phi_q - \frac{\pi}{2})} - D_N(f_q + \nu_k) e^{-i(\phi_q + \frac{\pi}{2})} \right) \right|^2 \\ &= \frac{N}{2S_E(\nu_k)} \left| \sum_{p=1}^{N_s} \alpha_q \left[\frac{\sin(N\pi(f_q - \nu_k))}{N \sin(\pi(f_q - \nu_k))} e^{i[(N+1)\pi(f_q - \nu_k) + (\phi_q - \frac{\pi}{2})]} - \frac{\sin(N\pi(f_q + \nu_k))}{N \sin(\pi(f_q + \nu_k))} e^{-i[(N+1)\pi(f_q + \nu_k) + (\phi_q + \frac{\pi}{2})]} \right] \right|^2 \\ &= \frac{N}{2S_E(\nu_k)} \left| \sum_{p=1}^{N_s} \alpha_q \left(\kappa_1(p) e^{i\theta_1(p)} - \kappa_2(p) e^{i\theta_2(p)} \right) \right|^2 \\ &= \frac{N}{2S_E(\nu_k)} \left| \sum_{p=1}^{N_s} \alpha_q \left(z_1(p) - z_2(p) \right) \right|^2\end{aligned}\tag{B.10}$$

with the complex numbers $z_1(p)$ and $z_2(p)$ such as $|\kappa_1(p)| = |D_N(f_q - \nu_k)|$ and $|\kappa_2(p)| = |D_N(f_q + \nu_k)|$:

$$\begin{cases} z_1(p) = \kappa_1(p) e^{i\theta_1(p)} = \kappa_1(p) \cos(\theta_1(p)) + i\kappa_1(p) \sin(\theta_1(p)) = a_1(p) + ib_1(p) \\ z_2(p) = \kappa_2(p) e^{i\theta_2(p)} = \kappa_2(p) \cos(\theta_2(p)) + i\kappa_2(p) \sin(\theta_2(p)) = a_2(p) + ib_2(p) \end{cases}$$

and the associated complex parameters in the geometric form (module and argument):

$$\begin{cases} \kappa_1(p) = \frac{\sin(N\pi(f_q - \nu_k))}{N \sin(\pi(f_q - \nu_k))} \\ \kappa_2(p) = \frac{\sin(N\pi(f_q + \nu_k))}{N \sin(\pi(f_q + \nu_k))} \\ \theta_1(p) = +[(N+1)\pi(f_q - \nu_k) + (\phi_q - \frac{\pi}{2})] \\ \theta_2(p) = -[(N+1)\pi(f_q + \nu_k) + (\phi_q + \frac{\pi}{2})]. \end{cases}$$

Then, we can develop the last line of (B.10) to obtain:

$$\lambda_k = \frac{N}{2S_E(\nu_k)} \left| \sum_{p=1}^{N_s} (\alpha_q \kappa_q) e^{i\theta_q} \right|^2 = \frac{N}{2S_E(\nu_k)} \left| \sum_{p=1}^{N_s} z_q \right|^2$$

with $z_q = z(p)$:

$$\begin{aligned} z_q &= \alpha_q [z_1(p) - z_2(p)] \\ &= \alpha_q [a_1(p) + ib_1(p) - (a_2(p) + ib_2(p))] \\ &= \alpha_q [(a_1(p) - a_2(p)) + i(b_1(p) - b_2(p))] \\ &= \alpha_q \kappa_q e^{i\theta_q} \end{aligned} \tag{B.11}$$

If we now develop the two new terms (κ_q and θ_q , we avoid in the following the indice (p) to be more clear):

$$\begin{aligned} \kappa_q &= |z_1(p) - z_2(p)| \\ &= \sqrt{(a_1(p) - a_2(p))^2 + (b_1(p) - b_2(p))^2} \\ &= \sqrt{(\kappa_1 \cos(\theta_1) - \kappa_2 \cos(\theta_2))^2 + (\kappa_1 \sin(\theta_1) - \kappa_2 \sin(\theta_2))^2} \\ &= \sqrt{(\kappa_1 \cos(\theta_1))^2 + (\kappa_2 \cos(\theta_2))^2 - 2\kappa_1 \kappa_2 \cos(\theta_1) \cos(\theta_2) + (\kappa_1 \sin(\theta_1))^2 + (\kappa_2 \sin(\theta_2))^2 - 2\kappa_1 \kappa_2 \sin(\theta_1) \sin(\theta_2)} \\ &= \sqrt{\kappa_1^2 (\cos(\theta_1)^2 + \sin(\theta_1)^2) + \kappa_2^2 (\cos(\theta_2)^2 + \sin(\theta_2)^2) - 2\kappa_1 \kappa_2 (\cos(\theta_1) \cos(\theta_2) + \sin(\theta_1) \sin(\theta_2))} \\ &= \sqrt{\kappa_1^2 + \kappa_2^2 - 2\kappa_1 \kappa_2 \cos(\theta_1 - \theta_2)} \\ &= \sqrt{K_N(f_q - \nu_k) + K_N(f_q + \nu_k) - 2 \frac{\sin(N\pi(f_q - \nu_k)) \sin(N\pi(f_q + \nu_k))}{N \sin(\pi(f_q - \nu_k)) N \sin(\pi(f_q + \nu_k))} \cos(\theta_1 - \theta_2)} \\ &= \sqrt{K_N(f_q - \nu_k) + K_N(f_q + \nu_k) - 2 \frac{\sin(N\pi(f_q - \nu_k)) \sin(N\pi(f_q + \nu_k))}{N \sin(\pi(f_q - \nu_k)) N \sin(\pi(f_q + \nu_k))} \cos(2\pi(N+1)f_q + 2\phi_q)} \end{aligned}$$

as

$$\theta_1 - \theta_2 = +[(N+1)\pi(f_q - \nu_k) + (\phi_q - \frac{\pi}{2})] - (-)[(N+1)\pi(f_q + \nu_k) + (\phi_q + \frac{\pi}{2})] = 2\pi(N+1)f_q + 2\phi_q$$

Now, for the argument of (B.11):

$$\theta_q = \angle z_q = \begin{cases} \arctan\left(\frac{b_1(p) - b_2(p)}{a_1(p) - a_2(p)}\right) \bmod [2\pi] & \text{if } (a_1 - a_2) > 0 \\ \arctan\left(\frac{b_1(p) - b_2(p)}{a_1(p) - a_2(p)}\right) + \pi \bmod [2\pi] & \text{if } (a_1 - a_2) < 0 \text{ and } (b_1 - b_2) > 0 \\ \arctan\left(\frac{b_1(p) - b_2(p)}{a_1(p) - a_2(p)}\right) - \pi \bmod [2\pi] & \text{if } (a_1 - a_2) < 0 \text{ and } (b_1 - b_2) < 0 \end{cases}$$

Finally, we use the fact that, for all complex $z_j = \kappa_j e^{i\theta_j}$, we have the relation

$$\left| \sum_j z_j \right|^2 = \sum_j \left[(\kappa_j)^2 + 2 \sum_{\ell=j+1} \kappa_j \kappa_\ell \cos(\theta_j - \theta_\ell) \right]$$

as from [Kumar, 2010]:

$$\begin{aligned} \left(\sum_j z_j \right) \left(\sum_\ell \bar{z}_\ell \right) &= \sum_j |z_j|^2 + \sum_{\ell > j} z_j \bar{z}_\ell \\ &= \sum_j \kappa_j^2 + \sum_{\ell > j} \kappa_j \kappa_\ell e^{i(\theta_j - \theta_\ell)} \\ &= \sum_j \kappa_j^2 + 2 \sum_{\ell=j+1} \kappa_j \kappa_\ell \cos(\theta_j - \theta_\ell) \end{aligned} \tag{B.12}$$

as the cosine terms in the exponential double up since cosine is even:

$$\cos(\theta_j - \theta_\ell) + \cos(\theta_\ell - \theta_j) = 2 \cos(\theta_j - \theta_\ell)$$

and the sine terms cancel since sine is odd:

$$\sin(\theta_j - \theta_\ell) + \sin(\theta_\ell - \theta_j) = 0.$$

So, we replace κ_j by $\alpha_q \kappa_q$ to obtain the final expression for the non centrality parameter given in (3.6):

$$\begin{aligned} \lambda_k &= \frac{N}{2S_E(\nu_k)} \left| \sum_{p=1}^{N_s} z_q \right|^2 \\ &= \frac{N}{2S_E(\nu_k)} \sum_{q=1}^{N_s} \left[\alpha_q^2 \kappa_q^2 + 2\alpha_q \kappa_q \sum_{\ell=q+1}^{N_s} \alpha_\ell \kappa_\ell \cos(\theta_q - \theta_\ell) \right]. \end{aligned} \tag{B.13}$$

Note that if all signal frequencies $\{f_q\}$ fall on the Fourier frequency grid, the crossed term in (B.13) vanish owing to the orthogonality of the Fejér kernels centered at different signal frequencies. In this case, expression (B.13) precisely reduces to expression given in Remark 6.6 of [Li, 2014].

The non-centrality parameter λ_k is evaluated at the Fourier frequencies ν_k and corresponds to the signal power normalised by the noise power in the k^{th} frequency bin. If there is only noise in the data $\lambda_k = 0$ but if some signals are present $\lambda_k \neq 0$ for the frequencies affected by the signals. The expression (B.13) is composed of two sums depending on the signal parameters (α_q, f_q, φ_o) and to the observational parameters ($N, \Delta t$). The truncation of the time sequence and the presence of off-grid frequency signals cause the spectral leakage characterised by the Fejér kernel K_N (or spectral window) in the κ_q terms. This kernel is composed on a main central lobe and sidelobes with width depending on the spectral resolution ($\frac{1}{N\Delta t}$) and to the signal frequencies position on the Fourier grid. The magnitude of these sidelobes at neighbouring frequencies can be significant when the signal is off-grid and the superposition of all the “signals kernel sidelobes” are evaluated by the second sum in (B.13). The second sum tends to zero in case of signal frequencies on the Fourier grid as $K_N(0) = 1$ and so $\kappa_q(f_q) \rightarrow 1$ and $\kappa_q(\nu_k \neq f_q) \rightarrow 0$ (i.e giving, $\forall \nu_k$, the product $\kappa_q \kappa_\ell \rightarrow 0$).

Appendix C

Theoretical AR autocorrelation function

This appendix aims to derive the theoretical AR autocorrelation function r_E up to a number of considered lags (N_{lags}) for $\sigma_{AR}^2 = 1$. The main steps of the used algorithm (cf. p.65 of [Tingyan, 2010]) are summarized below.

1. Find the first terms of the autocorrelation function $\{r_{E,1}, \dots, r_{E,o}\}$ by solving linear system setting up the matrix :

- We evaluate the Toeplitz matrix T of size $(o+1) \times (o+1)$:

$$\mathbf{T} = \begin{bmatrix} 1 & 2 & \dots & o+1 \\ 2 & 1 & \dots & o \\ \dots & \dots & \dots & \dots \\ p+1 & p-1 & \dots & 1 \end{bmatrix}$$

- We define a matrice \mathbf{R} of size $(o+1) \times (o+1)$ such as :

$$R(i, T(i, j)) = R(i, T(i, j)) - c_j, \quad \text{for } i, j = 1, \dots, o+1.$$

- Then, the function evaluates the right-hand-side column vector of the Toeplitz matrice : ψ with size $(o+1) \times (1)$:

$$\psi(1) = 1, \quad \text{and } \psi(i) = \begin{pmatrix} \psi(1) \\ \psi(2) \\ \dots \\ \psi(i-1) \end{pmatrix} \times \begin{pmatrix} c_i \\ c_{i-1} \\ \dots \\ c_2 \end{pmatrix}, \quad \text{for } i = 2, \dots, o+1.$$

- Finally, the ACF of the first terms are given by solving the system of linear equations

$$\mathbf{R} \times \mathbf{r_E} = \psi,$$

with the MATLAB function " $r_E = mldivide(R, \psi)$ ".

2. The other autocorrelation coefficients $\{r_{E,o+1}, \dots, r_{E,N_{lags}}\}$ are given by iterations. For $i = o+1, \dots, N_{lags}$:

$$\mathbf{r_{E,i}} = \begin{pmatrix} c_1 \\ c_2 \\ \dots \\ c_o \end{pmatrix} \times \begin{pmatrix} r_{E,i} \\ r_{E,i-1} \\ \dots \\ r_{E,i-o+1} \end{pmatrix}.$$

3. Finally, the function normalizes the ACF coefficients :

$$\mathbf{r_E} = \frac{\mathbf{r_E}}{r_{E,1}}.$$

Appendix D

Generalized extreme value : proofs of equations

This appendix aims to prove the equations derived in Sec. 4.3. All the demonstrations can be found in [Coles, 2001].

Return level function (4.16)

In this part, we derive the return level expressions given in (4.16).

First, we have the GEV cdf (4.15):

$$G(z; \boldsymbol{\theta}_{GEV}) := \begin{cases} \exp\left\{-\left[1 + \xi\left(\frac{z - \mu}{\sigma}\right)\right]_+^{-1/\xi}\right\}, & \text{for } \xi \neq 0 \\ \exp\left\{-\exp\left\{-\left(\frac{z - \mu}{\sigma}\right)\right\}\right\}, & \text{for } \xi = 0, \end{cases}$$

The return level $\gamma(q; \boldsymbol{\theta}_{GEV})$ is associated to a return period q such as:

$$\Pr(z > \gamma(q; \boldsymbol{\theta}_{GEV})) = 1 - G(z \leq \gamma(q; \boldsymbol{\theta}_{GEV})) := q.$$

Using the inverse function of G , we obtain:

$$\gamma(q; \boldsymbol{\theta}_{GEV}) = G^{-1}(1 - q; \boldsymbol{\theta}_{GEV}). \quad (\text{D.1})$$

Let us define $u := 1 - q$ and isolate u in $y := G(u; \boldsymbol{\theta}_{GEV})$.

For $\xi \neq 0$ we have:

$$\begin{aligned} y &:= \exp\left\{-\left[1 + \xi\left(\frac{u - \mu}{\sigma}\right)\right]_+^{-1/\xi}\right\} \\ \log(y) &= -\left[1 + \xi\left(\frac{u - \mu}{\sigma}\right)\right]_+^{-1/\xi} \\ \left(-\log(y)\right)^{-\xi} &= 1 + \xi\left(\frac{u - \mu}{\sigma}\right) \\ u &= \mu - \frac{\sigma}{\xi}\left(1 - \left(-\log(y)\right)^{-\xi}\right). \end{aligned} \quad (\text{D.2})$$

For $\xi = 0$, we have:

$$\begin{aligned} y &:= \exp\left\{-\exp\left\{-\left(\frac{u - \mu}{\sigma}\right)\right\}\right\} \\ -\log\left(\log(y)\right) &= \frac{u - \mu}{\sigma} \\ u &= \mu - \sigma\left(\log\left(\log(y)\right)\right). \end{aligned} \quad (\text{D.3})$$

We replace u into $y = G(u)$ and y into u in (D.2) and (D.3) to obtain the inverse of G in order to evaluate the return level function by (D.1). Let us define $y_q := -\log(1 - q)$ in order to obtain (4.16):

$$\gamma(q; \boldsymbol{\theta}_{GEV}) = G^{-1}(1 - q; \boldsymbol{\theta}_{GEV}) = \begin{cases} \mu - \frac{\sigma}{\xi} (1 - y_q^{-\xi}), & \text{for } \xi \neq 0, \\ \mu - \sigma \log y_q, & \text{for } \xi = 0. \quad \square \end{cases}$$

Likelihood function (4.17)

In this part, we derive the log-likelihood function associated to the GEV distribution given in (4.17).

Assuming the independence of the b maxima $\mathbf{z} := [z_1, \dots, z_b]^\top$, the likelihood function can be written as:

$$\mathcal{L}(\mathbf{z}; \boldsymbol{\theta}_{GEV}) = \prod_{i=1}^b g(z_i; \boldsymbol{\theta}_{GEV}) = \prod_{i=1}^b \frac{dG(z_i; \boldsymbol{\theta}_{GEV})}{dz_i}, \quad (\text{D.4})$$

with g the GEV pdf and G the GEV cdf given in (4.15).

We first obtain g for the case $\xi \neq 0$ and then turn to the case $\xi = 0$.

- Case $\xi \neq 0$:

$$\begin{aligned} g(z; \boldsymbol{\theta}_{GEV}) &= \frac{d \exp\{u(z)\}}{dz} \\ &= \frac{du(z)}{dz} \exp\{u(z)\} \\ &= \left(\frac{1}{\sigma}\right) \left[1 + \xi \left(\frac{z - \mu}{\sigma}\right)\right]_+^{-\left(\frac{1}{\xi} + 1\right)} \exp\left\{-\left[1 + \xi \left(\frac{z - \mu}{\sigma}\right)\right]_+^{-1/\xi}\right\}, \end{aligned} \quad (\text{D.5})$$

$$\text{with } \begin{cases} u(z) = -\left[1 + \xi \left(\frac{z - \mu}{\sigma}\right)\right]_+^{-1/\xi}, \\ \frac{du(z)}{dz} = -\left(-\frac{1}{\xi}\right) \left(\frac{\xi}{\sigma}\right) \left[1 + \xi \left(\frac{z - \mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi} - 1} = \left(\frac{1}{\sigma}\right) \left[1 + \xi \left(\frac{z - \mu}{\sigma}\right)\right]_+^{-\left(\frac{1}{\xi} + 1\right)}. \end{cases}$$

- Case $\xi = 0$:

$$\begin{aligned} g(z; \boldsymbol{\theta}_{GEV}) &= \frac{d \exp\{-\exp[u(z)]\}}{dz} \\ &= -\frac{d \exp\{u(z)\}}{dz} \exp\{-\exp[u(z)]\} \\ &= -\frac{du(z)}{dz} \exp\{u(z)\} \exp\{-\exp[u(z)]\} \\ &= \left(\frac{1}{\sigma}\right) \exp\left\{-\frac{z - \mu}{\sigma}\right\} \exp\left\{-\exp\left[-\frac{z - \mu}{\sigma}\right]\right\}, \end{aligned} \quad (\text{D.6})$$

$$\text{with } \begin{cases} u(z) = -\left(\frac{z - \mu}{\sigma}\right), \\ \frac{du(z)}{dz} = -\frac{1}{\sigma}. \end{cases}$$

Consequently, with (D.4), (D.5), and (D.6), the log-likelihood function in (D.4) writes as given in (4.17):

$$\begin{aligned}
\ell(\mathbf{z}; \boldsymbol{\theta}_{GEV}) &:= \log \mathcal{L}(\mathbf{z}; \boldsymbol{\theta}_{GEV}) \\
&= \log \left(\prod_{i=1}^b g(z_i; \boldsymbol{\theta}_{GEV}) \right) \\
&= \sum_{i=1}^b \log g(z_i; \boldsymbol{\theta}_{GEV}) \\
&= \begin{cases} \sum_{i=1}^b \log \left(\frac{1}{\sigma} \right) + \sum_{i=1}^b \log \left(\left[1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) \right]_+^{-\left(\frac{1}{\xi} + 1 \right)} \right) + \sum_{i=1}^b \log \left(\exp \left\{ - \left[1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) \right]_+^{-1/\xi} \right\} \right) & \text{for } \xi \neq 0, \\ \sum_{i=1}^b \log \left(\frac{1}{\sigma} \right) + \sum_{i=1}^b \log \left(\exp \left\{ - \left(\frac{z_i - \mu}{\sigma} \right) \right\} \right) + \sum_{i=1}^b \log \left(\exp \left\{ - \exp \left[- \frac{z_i - \mu}{\sigma} \right] \right\} \right) & \text{for } \xi = 0 \end{cases} \\
&= \begin{cases} -b \log(\sigma) - \left(1 + \frac{1}{\xi} \right) \sum_{i=1}^b \log \left(\left[1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) \right]_+ \right) - \sum_{i=1}^b \left[1 + \xi \frac{z_i - \mu}{\sigma} \right]_+^{-1/\xi}, & \text{for } \xi \neq 0, \\ -b \log(\sigma) - \sum_{i=1}^b \left(\frac{z_i - \mu}{\sigma} \right) - \sum_{i=1}^b \exp \left\{ - \left(\frac{z_i - \mu}{\sigma} \right) \right\}, & \text{for } \xi = 0. \quad \square \end{cases}
\end{aligned} \tag{D.7}$$

The log-likelihood system of equation

In this part, in order to estimate the GEV parameters, we derive the likelihood system of equation associated to the maximization of the likelihood given in (D.4). As it is more convenient to work with the log-likelihood function ℓ (4.17), we derive the log-likelihood with respect to the three GEV parameters (ξ, μ, σ) :

$$\begin{cases} \frac{d \ell(z; \xi, \mu, \sigma)}{d\xi} = 0, \\ \frac{d \ell(z; \xi, \mu, \sigma)}{d\sigma} = 0, \\ \frac{d \ell(z; \xi, \mu, \sigma)}{d\mu} = 0. \end{cases} \tag{D.8}$$

Derivation of (4.17) with respect to ξ

- First case, $\xi \neq 0$:

$$\frac{d \ell(z_i; \xi, \mu, \sigma)}{d\xi} = 0 - \frac{d \{f(\xi) \times h(\xi)\}}{d\xi} - \sum_{i=1}^b \frac{d \{u(\xi)^{v(\xi)}\}}{d\xi}, \tag{D.9}$$

$$\text{with } \begin{cases} f(\xi) = \left(1 + \frac{1}{\xi}\right) & \Rightarrow \frac{df(\xi)}{d\xi} = -\frac{1}{\xi^2}, \\ h(\xi) = \sum_{i=1}^b \log\left(1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right) & \Rightarrow \frac{dh(\xi)}{d\xi} = \sum_{i=1}^b \left(\frac{z_i - \mu}{\sigma}\right) \left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right]^{-1}, \\ u(\xi) = \left(1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right) & \Rightarrow \frac{du(\xi)}{d\xi} = \left(\frac{z_i - \mu}{\sigma}\right), \\ v(\xi) = -\frac{1}{\xi} & \Rightarrow \frac{dv(\xi)}{d\xi} = \frac{1}{\xi^2}. \end{cases}$$

So we can deduce the two terms of (D.9):

– First term

$$\begin{aligned} \frac{d\{f(\xi) \times h(\xi)\}}{d\xi} &= \frac{df(\xi)}{d\xi} \times h(\xi) + \frac{dh(\xi)}{d\xi} \times f(\xi) \\ &= -\left(\frac{1}{\xi^2}\right) \sum_{i=1}^b \log\left(1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right) + \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^b \left(\frac{z_i - \mu}{\sigma}\right) \left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right]^{-1}. \end{aligned}$$

– Second term

$$\begin{aligned} \frac{d\{u(\xi)^{v(\xi)}\}}{d\xi} &= u(\xi)^{v(\xi)} \times \left(\ln(u(\xi)) \times \frac{dv(\xi)}{d\xi} + v(\xi) \times \frac{1}{u(\xi)} \frac{du(\xi)}{d\xi}\right) \\ &= \left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}} \left(\frac{1}{\xi^2} \ln\left\{\left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right]_+\right\} - \frac{1}{\xi} \left(\frac{z_i - \mu}{\sigma}\right) \left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right]_+^{-1}\right). \end{aligned}$$

So (D.9) writes:

$$\begin{aligned} \frac{d\ell(\xi, \mu, \sigma; z_i)}{d\xi} &= -\frac{d\{f(\xi) \times h(\xi)\}}{d\xi} - \sum_{i=1}^b \frac{d\{u(\xi)^{v(\xi)}\}}{d\xi}, \\ &= +\left(\frac{1}{\xi^2}\right) \sum_{i=1}^b \log\left(1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right) \\ &\quad - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^b \left(\frac{z_i - \mu}{\sigma}\right) \left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right]^{-1} \tag{D.10} \\ &\quad - \sum_{i=1}^b \left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}} \left(\frac{1}{\xi^2} \ln\left\{\left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right]_+\right\} - \frac{1}{\xi} \left(\frac{z_i - \mu}{\sigma}\right) \left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right]_+^{-1}\right). \end{aligned}$$

- Second case, $\xi = 0$: For $\xi = 0$, the likelihood is not defined.

Derivation of (4.17) with respect to μ

- First case, $\xi \neq 0$:

$$\frac{d\ell(z; \xi, \mu, \sigma)}{d\mu} = 0 - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^b \frac{d\log\{u(\mu)\}}{d\mu} - \sum_{i=1}^b \frac{d u(\mu)^{-\frac{1}{\xi}}}{d\mu} \tag{D.11}$$

with

$$u(\mu) = 1 + \xi\left(\frac{z_i - \mu}{\sigma}\right) \quad \Rightarrow \quad \frac{du(\mu)}{d\mu} = -\left(\frac{\xi}{\sigma}\right).$$

So (D.11) is finally:

$$\begin{aligned}\frac{d \ell(z; \xi, \mu, \sigma)}{d\mu} &= -\left(1 + \frac{1}{\xi}\right) \sum_{i=1}^b \left(-\frac{\xi}{\sigma}\right) \left[1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right]_+^{-1} - \sum_{i=1}^b \left(-\frac{1}{\xi}\right) \left(-\frac{\xi}{\sigma}\right) \left[1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right]_+^{-\left(1 + \frac{1}{\xi}\right)} \\ &= \frac{\xi + 1}{\sigma} \sum_{i=1}^b \left[1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right]_+^{-1} - \frac{1}{\sigma} \sum_{i=1}^b \left[1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right]_+^{-\left(1 + \frac{1}{\xi}\right)}.\end{aligned}\quad (\text{D.12})$$

- Second case, $\xi = 0$:

$$\begin{aligned}\frac{d \ell(z; \xi, \mu, \sigma)}{d\mu} &= 0 - \sum_{i=1}^b \left(-\frac{1}{\sigma}\right) - \sum_{i=1}^b (-1) \times (-1) \times \frac{1}{\sigma} \exp\left\{-\left(\frac{z_i - \mu}{\sigma}\right)\right\} \\ &= \frac{1}{\sigma} \left(1 - \exp\left\{-\left(\frac{z_i - \mu}{\sigma}\right)\right\}\right)\end{aligned}\quad (\text{D.13})$$

Derivation of (4.17) with respect to σ

- First case, $\xi \neq 0$:

$$\frac{d \ell(z; \xi, \mu, \sigma)}{d\sigma} = -b \frac{d \log\{\sigma\}}{d\sigma} - \left(1 + \frac{1}{\xi}\right) \times \sum_{i=1}^b \frac{d \log\{u(\sigma)\}}{d\sigma} - \sum_{i=1}^b \frac{d \log\{u(\sigma)^{-\frac{1}{\xi}}\}}{d\sigma} \quad (\text{D.14})$$

with

$$\begin{cases} \frac{d \log(\sigma)}{d\sigma} = -\frac{b}{\sigma}, \\ u(\sigma) = 1 + \xi \left(\frac{z_i - \mu}{\sigma}\right) \end{cases} \Rightarrow \frac{du(\sigma)}{d\sigma} = -\left(\frac{1}{\sigma^2}\right) \xi(z_i - \mu).$$

So (D.14) is finally:

$$\begin{aligned}\frac{d \ell(\xi, \mu, \sigma; z_i)}{d\sigma} &= -\frac{b}{\sigma} + \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^b \left[1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right]_+^{-1} \left(\frac{1}{\sigma^2}\right) \xi(z_i - \mu) - \frac{1}{\xi} \sum_{i=1}^b \xi \left(\frac{z_i - \mu}{\sigma^2}\right) \left[1 + \xi \left(\frac{z_i - \mu}{\sigma^2}\right)\right]_+^{-\left(1 + \frac{1}{\xi}\right)} \\ &= -\frac{b}{\sigma} + \frac{\xi + 1}{\sigma^2} \sum_{i=1}^b \frac{(z_i - \mu)}{\left[1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right]_+} - \frac{1}{\sigma^2} \sum_{i=1}^b \frac{(z_i - \mu)}{\left[1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right]_+^{1 + \frac{1}{\xi}}}\end{aligned}\quad (\text{D.15})$$

- Second case, $\xi = 0$:

$$\frac{d \ell(z; \xi, \mu, \sigma)}{d\sigma} = \frac{1}{\sigma} \left(-b + \sum_{i=1}^b \left(\frac{z_i - \mu}{\sigma}\right) - \frac{1}{\sigma} \sum_{i=1}^b \exp\left\{-\left(\frac{z_i - \mu}{\sigma}\right)\right\}\right) \quad (\text{D.16})$$

It does not exist any analytical solution to solve equations (D.10), (D.12), (D.13), (D.15), and (D.16). Consequently, these equations have to be solved iteratively (e.g. using the Newton-Raphson method) to obtain the MLE parameters $\hat{\boldsymbol{\theta}}_{GEV} := [\hat{\xi}, \hat{\mu}, \hat{\sigma}]^\top$.

Standard errors on the GEV parameters

In this part, we describe one common method to first, derive the standard errors on the estimated GEV parameters $\hat{\boldsymbol{\theta}}_{GEV} := [\hat{\xi}, \hat{\mu}, \hat{\sigma}]^\top$ and then the confidence interval for the return level function (4.16) (namely the ‘‘delta method’’). This method is based on the variance-covariance information matrix and is implementable only for $\xi > -0.5$. One can refer to Sec. 3.3.2, p. 55 of [Coles, 2001] for a more detailed description.

The variance-covariance matrix is obtained via inversion of the observed information matrix, which is defined as the negative of the second-order partial derivatives of the negative log-likelihood (i.e. the Hessian matrix). We avoid in the following the dependence in \mathbf{z} to be more clear.

- The information matrix of the log-likelihood is:

$$I_\ell(\boldsymbol{\theta}_{GEV}) := -\nabla\nabla^T \ell(\boldsymbol{\theta}_{GEV}) = \begin{pmatrix} -\frac{\partial^2 \ell(\boldsymbol{\theta}_{GEV})}{\partial \xi^2} & -\frac{\partial^2 \ell(\boldsymbol{\theta}_{GEV})}{\partial \xi \partial \mu} & -\frac{\partial^2 \ell(\boldsymbol{\theta}_{GEV})}{\partial \xi \partial \sigma} \\ -\frac{\partial^2 \ell(\boldsymbol{\theta}_{GEV})}{\partial \xi \partial \mu} & -\frac{\partial^2 \ell(\boldsymbol{\theta}_{GEV})}{\partial \mu^2} & -\frac{\partial^2 \ell(\boldsymbol{\theta}_{GEV})}{\partial \mu \partial \sigma} \\ -\frac{\partial^2 \ell(\boldsymbol{\theta}_{GEV})}{\partial \xi \partial \sigma} & -\frac{\partial^2 \ell(\boldsymbol{\theta}_{GEV})}{\partial \sigma \partial \mu} & -\frac{\partial^2 \ell(\boldsymbol{\theta}_{GEV})}{\partial \sigma^2} \end{pmatrix}$$

- The observed information matrix is : $\hat{I}_\ell(\hat{\boldsymbol{\theta}}_{GEV})$.
- The approximate variance-covariance matrix results from the inversion of \hat{I}_ℓ :

$$\hat{V}_\ell(\hat{\boldsymbol{\theta}}_{GEV}) = \hat{I}_\ell(\hat{\boldsymbol{\theta}}_{GEV})^{-1}.$$

The diagonal elements give the variance of the estimated parameters $\hat{\boldsymbol{\theta}}_{GEV}$ and the off-diagonal elements the covariance between two of them. Thanks to the asymptotic normality of the MLE for any shape > 0.5 [Smith, 1985], the square roots of the diagonal elements of \hat{V}_ℓ give the expected standard errors for $\hat{\boldsymbol{\theta}}_{GEV}$. For instance, for a 95% confidence interval we obtain :

$$\begin{cases} \hat{\xi} \pm 1.96\sqrt{\hat{V}_{\ell,11}(\hat{\xi})}, \\ \hat{\mu} \pm 1.96\sqrt{\hat{V}_{\ell,22}(\hat{\mu})}, \\ \hat{\sigma} \pm 1.96\sqrt{\hat{V}_{\ell,33}(\hat{\sigma})}. \end{cases} \quad (\text{D.17})$$

It follows that confidence interval for the return level function (4.16) are obtainable, which is very useful for hypothesis testing (for the purpose of choosing between the 3 types of extreme distributions). Using the delta method, the variance-covariance matrix of the return level function writes:

$$\hat{V}_{\hat{\gamma}}(q; \hat{\boldsymbol{\theta}}_{GEV}) \approx \nabla \hat{\gamma}(q; \hat{\boldsymbol{\theta}}_{GEV})^T \hat{V}_\ell(\hat{\boldsymbol{\theta}}_{GEV}) \nabla \hat{\gamma}(q; \hat{\boldsymbol{\theta}}_{GEV}), \quad (\text{D.18})$$

with $\nabla \hat{\gamma}(q; \hat{\boldsymbol{\theta}}_{GEV})$ such as:

$$\nabla \hat{\gamma}(q; \hat{\boldsymbol{\theta}}_{GEV}) := \begin{cases} \left[\frac{\partial \hat{\gamma}(q; \hat{\boldsymbol{\theta}}_{GEV})}{\partial \hat{\xi}}, \frac{\partial \hat{\gamma}(q; \hat{\boldsymbol{\theta}}_{GEV})}{\partial \hat{\mu}}, \frac{\partial \hat{\gamma}(q; \hat{\boldsymbol{\theta}}_{GEV})}{\partial \hat{\sigma}} \right]^\top = \left[\hat{\sigma} \frac{1 - y_q^{-\hat{\xi}} - \hat{\xi}^{-1} y_q^{-\hat{\xi}} \log(y_q)}{\hat{\xi}^2}, 1, -\frac{1 - y_q^{-\hat{\xi}}}{\hat{\xi}} \right]^\top, \\ \left[\frac{\partial \hat{\gamma}(q; \hat{\boldsymbol{\theta}}_{GEV})}{\partial \hat{\mu}}, \frac{\partial \hat{\gamma}(q; \hat{\boldsymbol{\theta}}_{GEV})}{\partial \hat{\sigma}} \right]^\top = [1, -\log(y_q)]^\top, \quad \text{for } \hat{\xi} = 0. \end{cases} \quad (\text{D.19})$$

$\nabla \hat{\gamma}(q; \hat{\boldsymbol{\theta}}_{GEV})$ is evaluated at the MLE of the considered GEV model $\hat{\boldsymbol{\theta}}_{GEV}$. The 95% confidence interval for $\hat{\gamma}(q; \hat{\boldsymbol{\theta}}_{GEV})$ is then derived by

$$\hat{\gamma}(q; \hat{\boldsymbol{\theta}}_{GEV}) \pm 1.96\sqrt{\hat{V}_{\hat{\gamma}}(q; \hat{\boldsymbol{\theta}}_{GEV})}. \quad (\text{D.20})$$

One notes that other methods exist to estimate the confidence interval of the MLE parameters and the return level function such as parametric/non parametric bootstrap and profile likelihood for example (cf. [Coles, 2001]).

Bibliography

- [Abramowitz et al., 1972] Abramowitz, M. et al. (1972). *Spectral Analysis and Time Series*. Dover Publications.
- [Aigrain et al., 2012] Aigrain, S. et al. (2012). A simple method to estimate radial velocity variations due to stellar activity using photometry. *MNRAS*, 419:3147–3158.
- [Aittokallio et al., 2001] Aittokallio, T. et al. (2001). Testing for Periodicity in Signals: An Application to Detect Partial Upper Airway Obstruction during Sleep. *Journal of Theoretical Medicine*, 3(4):231–245.
- [Akaike, 1969] Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21(1):243–247.
- [Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *Transactions on Automatic Control*, 19(6):716–723.
- [Aldor-Noiman et al., 2013] Aldor-Noiman, S. et al. (2013). The power to see: A new graphical test of normality. *The American Statistician*, 67(4):249–260.
- [Allende Prieto, 2007] Allende Prieto, C. (2007). Velocities from Cross-Correlation: A Guide for Self-Improvement. *The Astronomical Journal*, 134:1843–1848.
- [Anderson, 1971] Anderson, T. (1971). *The Statistical Analysis of Time Series*. New York: John Wiley & Sons.
- [Anderson and Darling, 1954] Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769.
- [Anglada-Escudé et al., 2013] Anglada-Escudé, G. et al. (2013). A dynamically-packed planetary system around GJ 667C with three super-Earths in its habitable zone. *Astronomy and Astrophysics*, 556:A126.
- [Anglada-Escudé et al., 2016] Anglada-Escudé, G. et al. (2016). A terrestrial planet candidate in a temperate orbit around Proxima Centauri. *Nature*, 536:437–440.
- [Arias-Castro et al., 2011] Arias-Castro, E. et al. (2011). Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *Annals of Statistics*, 39(5):2533–2556.
- [Artis et al., 2004] Artis, M. et al. (2004). The detection of hidden periodicities: A comparison of alternative methods. Economics Working Papers ECO2004/10, European University Institute.
- [Asplund et al., 1999] Asplund, M. et al. (1999). 3D hydrodynamical model atmospheres of metal-poor stars. Evidence for a low primordial Li abundance. *Astronomy and Astrophysics*, 346:L17–L20.

- [Asplund et al., 2004] Asplund, M., Grevesse, N., Sauval, A. J., Allende Prieto, C., and Kiselman, D. (2004). Line formation in solar granulation. IV. [O I], O I and OH lines and the photospheric O abundance. *Astronomy and Astrophysics*, 417:751–768.
- [Asplund et al., 2009] Asplund, M., Grevesse, N., Sauval, A. J., and Scott, P. (2009). The Chemical Composition of the Sun. *Annual Review of Astronomy and Astrophysics*, 47:481–522.
- [Asplund et al., 2000] Asplund, M., Nordlund, Å., Trampedach, R., Allende Prieto, C., and Stein, R. F. (2000). Line formation in solar granulation. I. Fe line shapes, shifts and asymmetries. *Astronomy and Astrophysics*, 359:729–742.
- [Astudillo-Defru, 2015] Astudillo-Defru, N. (2015). *Recherche de planetes habitables autour de naines M*. PhD thesis, Universite Grenoble Alpes.
- [Baghi et al., 2016] Baghi, Q., Métris, G., Bergé, J., Christophe, B., Touboul, P., and Rodrigues, M. (2016). Gaussian regression and power spectral density estimation with missing data: The MICROSCOPE space mission as a case study. *Physical Review D*, 93(12):122007.
- [Baglin et al., 2006] Baglin, A. et al. (2006). Scientific Objectives for a Minisat: CoRoT. In Fridlund, M., Baglin, A., Lochard, J., and Conroy, L., editors, *The CoRoT Mission Pre-Launch Status - Stellar Seismology and Planet Finding*, volume 1306 of *ESA Special Publication*, page 33.
- [Balan and Lahav, 2009] Balan, S. T. and Lahav, O. (2009). exofit: orbital parameters of extrasolar planets from radial velocities. *Monthly Notices of the Royal Astronomical Society*, 394(4):1936.
- [Baliunas et al., 1995] Baliunas, S. L. et al. (1995). Chromospheric variations in main-sequence stars. *The Astrophysical Journal*, 438:269–287.
- [Ballerini et al., 2012] Ballerini, P., Micela, G., Lanza, A. F., and Pagano, I. (2012). Multi-wavelength flux variations induced by stellar magnetic activity: effects on planetary transits. *Astronomy and Astrophysics*, 539:A140.
- [Baluev, 2008] Baluev, R. (2008). Assessing the statistical significance of periodogram peaks. *MNRAS*, 385:1279–1285.
- [Baluev, 2013a] Baluev, R. (2013a). Detecting multiple periodicities in observational data with the multifrequency periodogram - I. Analytic assessment of the statistical significance. *MNRAS*, 436:807–818.
- [Baluev, 2015] Baluev, R. (2015). Keplerian periodogram for Doppler exoplanet detection: optimized computation and analytic significance thresholds. *MNRAS*, 446:1478–1492.
- [Baluev, 2013b] Baluev, R. V. (2013b). PlanetPack: A radial-velocity time-series analysis tool facilitating exoplanets detection, characterization, and dynamical simulations. *Astronomy and Computing*, 2:18–26.
- [Baranne et al., 1996] Baranne, A. et al. (1996). ELODIE: A spectrograph for accurate radial velocity measurements. *Astronomy and Astrophysics*, 119:373–390.
- [Barnard, 1916] Barnard, E. E. (1916). A small star with large proper-motion. *The Astronomical Journal*, 29:181–183.
- [Barnes et al., 2009] Barnes, R. et al. (2009). The HD 40307 Planetary System: Super-Earths or Mini-Neptunes? *The Astrophysical Journal*, 695:1006–1011.

- [Bartlett, 1950] Bartlett, M. (1950). Periodogram analysis and continuous spectra. *Biometrika*, 37:1–16.
- [Bartlett, 1955] Bartlett, M. (1955). An introduction to stochastic processes. *Quarterly Journal of the Royal Meteorological Society*, 81(350):650.
- [Basri and Brown, 2006] Basri, G. and Brown, M. E. (2006). Planetesimals to Brown Dwarfs: What is a Planet? *Annual Review of Earth and Planetary Sciences*, 34:193–216.
- [Beaulieu et al., 2006] Beaulieu, J.-P. et al. (2006). Discovery of a cool planet of 5.5 Earth masses through gravitational microlensing. *Nature*, 439:437–440.
- [Beaulieu et al., 2013] Beaulieu, J.-P., Tisserand, P., and Batista, V. (2013). Space based microlensing planet searches. In *European Physical Journal Web of Conferences*, volume 47 of *European Physical Journal Web of Conferences*, page 15001.
- [Bedding, 2011] Bedding, T. (2011). Probing the cores of red giant stars using asteroseismology with NASA’s Kepler Mission. *AGU Fall Meeting Abstracts*.
- [Bedding, 2014] Bedding, T. (2014). Solar-like oscillations: An observational perspective. In Palte, P. L. and Esteban, C., editors, *Asteroseismology*, page 60.
- [Belkacem et al., 2011] Belkacem, K., Goupil, M. J., Dupret, M. A., Samadi, R., Baudin, F., Noels, A., and Mosser, B. (2011). The underlying physical meaning of the $\nu_{max} - \nu_c$ relation. *Astronomy and Astrophysics*, 530:A142.
- [Bennett, 2008] Bennett, D. P. (2008). *Detection of Extrasolar Planets by Gravitational Microlensing*. Mason, J. W.
- [Berk and Jones, 1979] Berk, R. and Jones, D. (1979). Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 47(1):47–59.
- [Beuzit et al., 2006] Beuzit, J.-L. et al. (2006). Direct detection of exoplanets. In *Protostars and Planets V, Conference, October 24–28, 2005, Hilton Waikoloa Village, Hawai’i. Proceedings of the PPV Conference*.
- [Bhansali, 1979] Bhansali, R. (1979). A mixed spectrum analysis of the lynx data. *Journal of the Royal Statistical Society: Series A*, 142(2):199–209.
- [Bigot et al., 2011] Bigot, L. et al. (2011). The diameter of the CoRoT target HD 49933. *Astronomy and Astrophysics*, 534(3).
- [Bigot and Thévenin, 2008] Bigot, L. and Thévenin, F. (2008). 3D hydrodynamical simulations of stellar surfaces : Applications to Gaia. In Charbonnel, C. et al., editors, *SF2A-2008*, page 3.
- [Blackman and Tukey, 1958] Blackman, R. and Tukey, J. (1958). The measurement of power spectra from the point of view of communications engineering. *Bell System Technical Journal*, 37(1):185–282.
- [Blomberg, 2012] Blomberg, N. (2012). Higher criticism testing for signal detection in rare and weak models. Master thesis,, KTH Royal Institute of Technology, Stockholm, Sweden.
- [Bloomfield, 2000] Bloomfield, P. (2000). *Fourier Analysis of Time Series*. Wiley-Interscience.
- [Boardman et al., 2002] Boardman, A. et al. (2002). A study on the optimum order of autoregressive models for heart rate variability. *Physiological Measurement*, 23(2):325.

- [Boisse et al., 2011] Boisse, I., Bouchy, F., Hébrard, G., Bonfils, X., Santos, N., and Vauclair, S. (2011). Disentangling between stellar activity and planetary signals. In Prasad Choudhary, D. and Strassmeier, K. G., editors, *Physics of Sun and Star Spots*, volume 273 of *IAU Symposium*, pages 281–285.
- [Bölviken, 1983a] Bölviken, E. (1983a). The distribution of certain rational functions of order statistics from exponential distributions. *Scandinavian Journal of Statistics*, 10(2):117–123.
- [Bölviken, 1983b] Bölviken, E. (1983b). New tests of significance in periodogram analysis. *Scandinavian Journal of Statistics*, 10(1):1–9.
- [Bonfils et al., 2015] Bonfils, X. et al. (2015). ExTrA: Exoplanets in transit and their atmospheres. In *Techniques and Instrumentation for Detection of Exoplanets VII*, volume 9605 of *Proceedings of the SPIE*, page 96051L.
- [Borucki et al., 2010] Borucki, W. J. et al. (2010). Kepler Planet-Detection Mission: Introduction and First Results. *Science*, 327:977.
- [Borucki and Summers, 1984] Borucki, W. J. and Summers, A. L. (1984). The photometric method of detecting other planetary systems. *Icarus*, 58:121–134.
- [Bos et al., 2002] Bos, R., de Waele, S., and Broersen, P. (2002). Autoregressive spectral estimation by application of the burg algorithm to irregularly sampled data. *IEEE Transactions on Instrumentation and Measurement*, 51:1289 – 1294.
- [Bouchy, 2015] Bouchy, F. (2015). ELODIE & SOPHIE spectrographs: 20 years of continuous improvements in radial velocities. In Boisse, I. et al., editors, *Proceedings of colloquium 'Twenty years of giant exoplanets' held at OHP, France*, pages 1–7.
- [Bouchy and Carrier, 2001] Bouchy, F. and Carrier, F. (2001). P-mode observations on α Cen A. *Astronomy and Astrophysics*, 374:L5–L8.
- [Bouchy et al., 2001] Bouchy, F., Pepe, F., and Queloz, D. (2001). Fundamental photon noise limit to radial velocity measurements. *Astronomy and Astrophysics*, 374:733–739.
- [Boumier and Dame, 1993] Boumier, P. and Dame, L. (1993). GOLF: A resonance spectrometer for the observation of solar oscillations. 1: Numerical model of the sodium cell response. *Experimental Astronomy*, 4:87–104.
- [Bourguignon, 2006] Bourguignon, S. (2006). *Analyse spectrale à haute résolution de signaux irrégulièrement échantillonnés : application à l'Astrophysique*. PhD thesis, Université Paul Sabatier - Toulouse II.
- [Bourguignon et al., 2007] Bourguignon, S., Carfantan, H., and Böhm, T. (2007). SparSpec: a new method for fitting multiple sinusoids with irregularly sampled data. *Astronomy and Astrophysics*, 462:379–387.
- [Bozza et al., 2016] Bozza, V., Mancini, L., and Sozzetti, A. (2016). *Methods of Detecting Exoplanets: 1st Advanced School on Exoplanetary Science*. Springer.
- [Bretthorst, 2003] Bretthorst, G. (2003). Frequency Estimation And Generalized Lomb-Scargle Periodograms. *Statistical Challenges in Astronomy*, pages 309–329.
- [Bretthorst, 1988] Bretthorst, G. L. (1988). Introduction. In *Bayesian Spectrum Analysis and Parameter Estimation*, pages 1–11. Springer.
- [Brillinger, 1981] Brillinger, D. (1981). *Time Series : Data Analysis and Theory*. Holden Day, San Francisco.

- [Brockwell and Davis, 1991] Brockwell, P. and Davis, R. (1991). *Time series : theory and methods*. Springer.
- [Brogi et al., 2017] Brogi, M., Line, M., Bean, J., Désert, J.-M., and Schwarz, H. (2017). A Framework to Combine Low- and High-resolution Spectroscopy for the Atmospheres of Transiting Exoplanets. *The Astrophysical Journal*, 839:L2.
- [Brown and Gilliland, 1994] Brown, T. M. and Gilliland, R. L. (1994). Asteroseismology. *Annual Review of Astronomy and Astrophysics*, 32:37–82.
- [Capon, 1969] Capon, J. (1969). High-resolution frequency-wavenumber spectrum analysis. In *Proceedings of the IEEE (Volume:57 , Issue: 8)*, pages 1408–1418.
- [Casertano et al., 2008] Casertano, S. et al. (2008). Double-blind test program for astrometric planet detection with Gaia. *Astronomy and Astrophysics*, 482:699–729.
- [Cassan, 2014] Cassan, A. (2014). *Extrasolar planets detections and statistics through gravitational microlensing*. Habilitation à diriger des recherches, Université Pierre et Marie Curie.
- [Cegla et al., 2015] Cegla, H. M. et al. (2015). Understanding Astrophysical Noise from Stellar Surface Magneto-Convection. In van Belle, G. T. and Harris, H. C., editors, *18th Cambridge Workshop on Cool Stars, Stellar Systems, and the Sun*, volume 18 of *Cambridge Workshop on Cool Stars, Stellar Systems, and the Sun*, pages 567–574.
- [Cegla et al., 2013] Cegla, H. M., Shelyag, S., Watson, C. A., and Mathioudakis, M. (2013). Stellar surface magneto-convection as a source of astrophysical noise. i. multi-component parameterization of absorption line profiles. *The Astrophysical Journal*, 763(2):95.
- [Chiavassa et al., 2015] Chiavassa, A. et al. (2015). New view on exoplanet transits. Transit of Venus described using three-dimensional solar atmosphere STAGGER-grid simulations. *Astronomy and Astrophysics*, 576:A13.
- [Chiavassa et al., 2017] Chiavassa, A. et al. (2017). Measuring stellar granulation during planet transits. *Astronomy and Astrophysics*, 597:A94.
- [Chiu, 1989] Chiu, S. (1989). Detecting periodic components in a white gaussian time series. *Journal of the Royal Statistical Society: Series B*, 51(2):249–259.
- [Chiu, 1990] Chiu, S. (1990). Peak-insensitive parametric spectrum estimation. *Stochastic Processes and their Applications*, 35:121–140.
- [Chou et al., 1991] Chou, D.-Y. et al. (1991). Power spectra of solar convection. *The Astrophysical Journal*, 372:314–320.
- [Christensen-Dalsgaard, 2014] Christensen-Dalsgaard, J. (2014). Asteroseismology of red giants. In Palle, P. L. and Esteban, C., editors, *Asteroseismology*, page 194.
- [Coles, 2001] Coles, S. G. (2001). *An Introduction to Statistical Modelling of Extreme Values*. Springer-Verlag, London.
- [Cumming, 2004] Cumming, A. (2004). Detectability of extrasolar planets in radial velocity surveys. *MNRAS*, 354:1165–1176.
- [Cumming et al., 1999] Cumming, A., Marcy, G. W., and Butler, R. P. (1999). The Lick Planet Search: Detectability and Mass Thresholds. *The Astrophysical Journal*, 526:890–915.
- [Daniell, 1946] Daniell, P. (1946). Discussion on symposium on autocorrelation in time series. *Journal of the Royal Statistical Society*, 8:88–90.

- [David and Nagaraja, 2003] David, H. and Nagaraja, H. (2003). *Order Statistics, Third Edition*. Wiley.
- [Dawes, 1864] Dawes, W. R. (1864). Results of some recent Observations of the Solar Surface, with Remarks. *MNRAS*, 24:161.
- [de Waele and Broersen, 2000] de Waele, S. and Broersen, P. (2000). The burg algorithm for segments. *IEEE Transactions on Signal Processing*, 48:2876 – 2880.
- [Debosscher et al., 2007] Debosscher, J. et al. (2007). Automated supervised classification of variable stars. I. Methodology. *Astronomy and Astrophysics*, 475:1159–1183.
- [Del Moro, 2004] Del Moro, D. (2004). Solar granulation properties derived from three different time series. *Astronomy and Astrophysics*, 428:1007–1015.
- [Deming et al., 2013] Deming, D. et al. (2013). Infrared Transmission Spectroscopy of the Exoplanets HD 209458b and XO-1b Using the Wide Field Camera-3 on the Hubble Space Telescope. *The Astrophysical Journal*, 774:95.
- [Desort et al., 2007] Desort, M. et al. (2007). Search for exoplanets with the radial-velocity technique: quantitative diagnostics of stellar activity. *Astronomy and Astrophysics*, 473:983–993.
- [Deubner and Gough, 1984] Deubner, F.-L. and Gough, D. (1984). Helioseismology: Oscillations as a Diagnostic of the Solar Interior. *Annual Reviews of Astronomy and Astrophysics*, 22:593–619.
- [Ding et al., 2015] Ding, J. et al. (2015). Bridging AIC and BIC: a new criterion for autoregression. *ArXiv e-prints*.
- [Donoho and Jin, 2004] Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*.
- [Doppler, 1842] Doppler, C. (1842). *Über das farbige Licht der Doppelsterne und einiger anderer Gestirne des Himmels*. Abhandlungen einer Privatgesellschaft in Böhmen. NF.5,2. Calve.
- [Dravins, 1987] Dravins, D. (1987). Stellar granulation. I - The observability of stellar photospheric convection. *Astronomy and Astrophysics*, 172:200–224.
- [Dravins, 1990] Dravins, D. (1990). Stellar granulation. In Wallerstein, G., editor, *Cool Stars, Stellar Systems, and the Sun*, volume 9 of *Astronomical Society of the Pacific Conference Series*, pages 27–38.
- [Dravins, 1999] Dravins, D. (1999). Stellar Surface Convection, Line Asymmetries, and Wavelength Shifts. In Hearnshaw, J. B. and Scarfe, C. D., editors, *IAU Colloq. 170: Precise Stellar Radial Velocities*, volume 185 of *Astronomical Society of the Pacific Conference Series*, page 268.
- [Dravins, 2008] Dravins, D. (2008). “Ultimate” information content in solar and stellar spectra. Photospheric line asymmetries and wavelength shifts. *The Astrophysical Journal*, 492:199–213.
- [Dravins and Lind, 1984] Dravins, D. and Lind, J. (1984). Observing Stellar Granulation (Keynote). In Keil, S. L., editor, *Small-Scale Dynamical Processes in Quiet Stellar Atmospheres*, page 414.

- [Dravins et al., 1981] Dravins, D., Lindegren, L., and Nordlund, A. (1981). Solar granulation - Influence of convection on spectral line asymmetries and wavelength shifts. *Astronomy and Astrophysics*, 96:345–364.
- [Dumusque, 2016] Dumusque, X. (2016). Radial velocity fitting challenge. I. Simulating the data set including realistic stellar radial-velocity signals. *Astronomy and Astrophysics*, 593:A5.
- [Dumusque et al., 2011] Dumusque, X. et al. (2011). Planetary detection limits taking into account stellar noise. I. Observational strategies to reduce stellar oscillation and granulation effects. *The Astrophysical Journal*, 525:A140.
- [Dumusque et al., 2012] Dumusque, X. et al. (2012). An Earth-mass planet orbiting α Centauri B. *Nature*, 491:207–211.
- [Dumusque et al., 2015] Dumusque, X. et al. (2015). HARPS-N Observes the Sun as a Star. *The Astrophysical Journal Letters*, 814:L21.
- [Dumusque et al., 2017] Dumusque, X. et al. (2017). Radial-velocity fitting challenge. II. First results of the analysis of the data set. *Astronomy and Astrophysics*, 598:A133.
- [Einstein, 1936] Einstein, A. (1936). Lens-Like Action of a Star by the Deviation of Light in the Gravitational Field. *Science*, 84:506–507.
- [Eyer and Bartholdi, 1999] Eyer, L. and Bartholdi, P. (1999). Variable stars: Which Nyquist frequency? *Astronomy and Astrophysics*, 135:1–3.
- [Feng et al., 2016] Feng, F. et al. (2016). A Goldilocks principle for modelling radial velocity noise. *MNRAS*, 461:2440–2452.
- [Feng et al., 2017] Feng, F., Tuomi, M., and Jones, H. R. A. (2017). Agatha: disentangling periodic signals from correlated noise in a periodogram framework. *MNRAS*, 470:4794–4814.
- [Ferraz-Mello, 1981] Ferraz-Mello, S. (1981). Estimation of Periods from Unequally Spaced Observations. *The Astronomical Journal*, 86:619.
- [Figueira et al., 2010] Figueira, P. et al. (2010). Evidence against the young hot-Jupiter around BD +20 1790. *Astronomy and Astrophysics*, 513:L8.
- [Fischer et al., 2014] Fischer, D. et al. (2014). Exoplanet Detection Techniques. *Protostars and Planets VI*, pages 715–737.
- [Fisher, 1929] Fisher, R. (1929). Tests of Significance in Harmonic Analysis. *Royal Society of London Proc. Series A*, 125:54–59.
- [Fisher and Tippetts, 1928] Fisher, R. A. and Tippetts, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Math. Proc. Cambridge Philo. Society*, 24:180–190.
- [Ford et al., 2011] Ford, E. B. et al. (2011). Transit Timing Observations from Kepler. I. Statistical Analysis of the First Four Months. *The Astrophysical Journal Supplement*, 197:2.
- [Ford and Gregory, 2007] Ford, E. B. and Gregory, P. C. (2007). Bayesian Model Selection and Extrasolar Planet Detection. In Babu, G. J. and Feigelson, E. D., editors, *Statistical Challenges in Modern Astronomy IV*, volume 371 of *Astronomical Society of the Pacific Conference Series*, page 189.
- [Frescura et al., 2008] Frescura, F. et al. (2008). Significance of periodogram peaks and a pulsation mode analysis of the beta cephei star v403 car. *MNRAS*, 388(4):1693.

- [Fressin et al., 2013] Fressin, F. et al. (2013). The False Positive Rate of Kepler and the Occurrence of Planets. *The Astrophysical Journal*, 766:81.
- [Freytag et al., 2012] Freytag, B. et al. (2012). Simulations of stellar convection with CO5BOLD. *Journal of Computational Physics*, 231:919–959.
- [Gabriel et al., 1995] Gabriel, A. H. et al. (1995). Global Oscillations at Low Frequency from the SOHO Mission (GOLF). *Solar Physics*, 162:61–99.
- [Galland et al., 2005] Galland, F. et al. (2005). Extrasolar planets and brown dwarfs around A-F type stars. I. Performances of radial velocity measurements, first analyses of variations. *Astronomy and Astrophysics*, 443:337–345.
- [Garcia et al., 2005] Garcia, R. et al. (2005). Global solar Doppler velocity determination with the GOLF/SoHO instrument. *Astronomy and Astrophysics*, 442:385–395.
- [García, 2015] García, R. A. (2015). Observational techniques to measure solar and stellar oscillations. In *EAS Publications Series*, volume 73 of *EAS Publications Series*, pages 193–259.
- [Gillon et al., 2017] Gillon, M. et al. (2017). Seven temperate terrestrial planets around the nearby ultracool dwarf star TRAPPIST-1. *Nature*, 542:456–460.
- [Gontscharuk et al., 2014] Gontscharuk, V. et al. (2014). The intermediates take it all: Asymptotics of higher criticism statistics and a powerful alternative based on equal local levels. *Biometrical Journal*, 57(1):159–180.
- [Gontscharuk et al., 2016] Gontscharuk, V. et al. (2016). Goodness of fit tests in terms of local levels with special emphasis on higher criticism tests. *Bernoulli*, 22(3):1331–1363.
- [Gould and Loeb, 1992] Gould, A. and Loeb, A. (1992). Discovering planetary systems through gravitational microlenses. *The Astrophysical Journal*, 396:104–114.
- [Graham et al., 2013] Graham, M. J. et al. (2013). A comparison of period finding algorithms. *MNRAS*, 434:3423–3444.
- [Gray, 2005] Gray, D. F. (2005). *The Observation and Analysis of Stellar Photospheres*. UK: Cambridge University Press.
- [Gregory, 2016] Gregory, P. (2016). An apodized kepler periodogram for separating planetary and stellar activity signals. *MNRAS*, 458(3):2604–2633.
- [Grenander and Rosenblatt, 1957] Grenander, U. and Rosenblatt, M. (1957). *Statistical analysis of stationary time series*. John Wiley and Sons.
- [Griffin and Griffin, 1973] Griffin, R. and Griffin, R. (1973). On the possibility of determining stellar radial velocities to 0.01 km s^{-1} . *MNRAS*, 162:243–253.
- [Gryca, 1998] Gryca, J. (1998). Detection of multiple sinusoids buried in noise via balanced model truncation. *International Instrumentation and Measurement Technology Conference, IEEE*, 2:1353–1358.
- [Guillot, 1999] Guillot, T. (1999). Interior of Giant Planets Inside and Outside the Solar System. *Science*, 286.
- [Gupta et al., 2001] Gupta, S. et al. (2001). UPSO three channel fast photometer. *Bulletin of the Astronomical Society of India*, 29:479–486.

- [Gutiérrez-Soto et al., 2009] Gutiérrez-Soto, J. et al. (2009). Low-amplitude variations detected by CoRoT in the B8IIIe star HD 175869. *Astronomy and Astrophysics*, 506:133–141.
- [Hall and Jin, 2010] Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Annals of Statistics*, 38(3):1686–1732.
- [Hannan, 1961] Hannan, E. (1961). Testing for a jump in the spectral function. *Journal of the Royal Statistical Society: Series B*, 23(2):394–404.
- [Hannan and Quinn, 1979] Hannan, E. and Quinn, B. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B*, 41(2):190–195.
- [Hara et al., 2017] Hara, N. C., Boué, G., Laskar, J., and Correia, A. C. M. (2017). Radial velocity data analysis with compressed sensing techniques. *MNRAS*, 464:1220–1246.
- [Harvey, 1985] Harvey, K. (1985). The relationship between coronal bright points as seen in He I Lambda 10830 and the evolution of the photospheric network magnetic fields. *Aust. J. of Physics*, 38:875–883.
- [Hatzes, 2013] Hatzes, A. (2013). The Radial Velocity Detection of Earth-mass Planets in the Presence of Activity Noise: The Case of α Centauri Bb. *The Astrophysical Journal*, 770:133.
- [Hatzes and Cochran, 1992] Hatzes, A. P. and Cochran, W. D. (1992). Spectrograph Requirements for Precise Radial Velocity Measurements. In Ulrich, M.-H., editor, *European Southern Observatory Conference and Workshop Proceedings*, volume 40 of *European Southern Observatory Conference and Workshop Proceedings*, page 275.
- [Hatzes and Cochran, 2000] Hatzes, A. P. and Cochran, W. D. (2000). The Radial Velocity and Spectral Line Bisector Variability of Polaris. *The Astrophysical Journal*, 120:979–989.
- [Haywood, 2015] Haywood, R. D. (2015). *Hide and Seek: Radial-Velocity Searches for Planets around Active Stars*. PhD thesis, University of St Andrews.
- [Hébrard et al., 2004] Hébrard, G. et al. (2004). Evaporation Rate of Hot Jupiters and Formation of Chthonian Planets. In Beaulieu, J., Lecavelier Des Etangs, A., and Terquem, C., editors, *Extrasolar Planets: Today and Tomorrow*, volume 321 of *Astronomical Society of the Pacific Conference Series*, page 203.
- [Herschel, 1801] Herschel, W. (1801). Observations Tending to Investigate the Nature of the Sun, in Order to Find the Causes or Symptoms of Its Variable Emission of Light and Heat; With Remarks on the Use That May Possibly Be Drawn from Solar Observations. *Philosophical Transactions of the Royal Society of London Series I*, 91:265–318.
- [Horne and Baliunas, 1986] Horne, J. H. and Baliunas, S. L. (1986). A prescription for period analysis of unevenly sampled time series. *The Astrophysical Journal*, 302:757–763.
- [Ingster et al., 2010] Ingster, Y. et al. (2010). Detection boundary in sparse regression. *Electronic Journal of Statistics*, 4:1476–1526.
- [Ingster et al., 1998] Ingster, Y. I., Sapatinas, T., and Suslina, I. A. (1998). Minimax detection of a signal for besov bodies and balls. *Problems of Information Transmission*, 34(1):56–68.
- [Jager and Wellner, 2007] Jager, L. and Wellner, J. A. (2007). Goodness-of-fit tests via phi-divergences. *Annals of Statistics*, 35(5):2018–2053.
- [Jenkins et al., 2014] Jenkins, J. S. et al. (2014). Improved signal detection algorithms for unevenly sampled data. Six signals in the radial velocity data for GJ876. *MNRAS*, 441:2253–2265.

- [Jenkinson, 1955] Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, 81:158–171.
- [Jin, 2004] Jin, J. (2004). *Detecting a target in very noisy data from multiple looks*, volume 45 of *Lecture Notes–Monograph Series*, pages 255–286. Institute of Mathematical Statistics.
- [Johnson et al., 1994] Johnson, N. et al. (1994). *Continuous Univariate Distributions*. A Wiley-Interscience Publication.
- [Kane et al., 2016] Kane, S. R. et al. (2016). Stellar Activity and Exclusion of the Outer Planet in the HD 99492 System. *The Astrophysical Journal Letters*, 820:L5.
- [Kaplan and Goldman, 2014] Kaplan, D. and Goldman, M. (2014). True equality (of pointwise sensitivity) at last: a dirichland othersternative to Kolmogorov-Smirnov inference on distributions. *Technical report*.
- [Katz, 2008] Katz, R. W. (2008). Background on extreme value theory with emphasis on climate applications. Short course on statistics of extremes in climate change, Michigan State University.
- [Kavalieris and E.J., 1994] Kavalieris, L. and E.J., H. (1994). Determining the number of terms in a trigonometric regression. *Journal of Time Series Analysis*, 15:613–625.
- [Kay, 1998] Kay, S. (1998). *Fundamentals of Statistical signal processing. Vol II : Detection theory*. Prentice-Hall, Inc.
- [Kay, 1999] Kay, S. (1999). Adaptive detection for unknown noise power spectral densities. *Signal Processing, IEEE Transactions on*, 47(1):10–21.
- [Kjeldsen and Bedding, 2011] Kjeldsen, H. and Bedding, T. R. (2011). Amplitudes of solar-like oscillations: a new scaling relation. *Astronomy and Astrophysics*, 529:L8.
- [Koen, 2015a] Koen, C. (2015a). The analysis of indexed astronomical time series - xi. the statistics of oversampled white noise periodograms. *MNRAS*, 449(1):1098–1105.
- [Koen, 2015b] Koen, C. (2015b). The analysis of indexed astronomical time series - xii. the statistics of oversampled fourier spectra of noise plus a single sinusoid. *MNRAS*, 453(2):1793–1798.
- [Kolmogorov, 1933] Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell’Istituto Italiano degli Attuari*, 4:83–91.
- [Kovacs et al., 2002] Kovacs, G. et al. (2002). A box-fitting algorithm in the search for periodic transits. *The Astrophysical Journal*, 391:369–377.
- [Kreiss and Paparoditis, 2003] Kreiss, J. and Paparoditis, E. (2003). Autoregressive-aided periodogram bootstrap for time series. *Annals of Statistics*, 31:1923–55.
- [Kuchner and Seager, 2005] Kuchner, M. J. and Seager, S. (2005). Extrasolar Carbon Planets. *ArXiv Astrophysics e-prints*.
- [Kumar, 2010] Kumar, V. (2010). *Algebra For Iit Jee*. McGraw-Hill Education (India) Pvt Limited.
- [Ledwina, 1994] Ledwina, T. (1994). Data-driven version of neyman’s smooth test of fit. *Journal of the American Statistical Association*, 89(427):1000–1005.

- [Lefebvre et al., 2008] Lefebvre, S. et al. (2008). Variations of the solar granulation motions with height using the GOLF/SoHO experiment. *Astronomy and Astrophysics*, 490:1143–1149.
- [Léger et al., 2004] Léger, A. et al. (2004). A new family of planets? “Ocean-Planets”. *Icarus*, 169:499–504.
- [Lehmann and Romano, 2005] Lehmann, E. and Romano, J. (2005). *Testing Statistical Hypotheses*. Springer Texts in Statistics.
- [Leighton, 1960] Leighton, R. B. (1960). Notes. In Thomas, R. N., editor, *Aerodynamic Phenomena in Stellar Atmospheres*, volume 12 of *IAU Symposium*, pages 321–325.
- [Li and Siegmund, 2015] Li, J. and Siegmund, D. (2015). Higher criticism: p -values and criticism. *Annals of Statistics*, 43(3):1323–1350.
- [Li, 2014] Li, T. (2014). *Time series with mixed spectra*. CRC Press.
- [Liavas et al., 1998] Liavas, A. et al. (1998). A periodogram-based method for the detection of steady-state visually evoked potentials. *IEEE Transactions on biomedical engineering*, 45(2):242–248.
- [Lissauer et al., 2011] Lissauer, J. J. et al. (2011). A closely packed system of low-mass, low-density planets transiting Kepler-11. *Nature*, 470:53–58.
- [Lomb, 1976] Lomb, N. (1976). Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science*, 39:447–462.
- [Lovis et al., 2006] Lovis, C. et al. (2006). The exoplanet hunter HARPS: unequalled accuracy and perspectives toward 1 cm.s^{-1} precision. In *SPIE*, volume 6269 of *Conference Series*, page 62690P.
- [Lu and Zimmerman, 2005] Lu, N. and Zimmerman, D. (2005). Testing for directional symmetry in spatial dependence using the periodogram. *Journal of Statistical Planning and Inference*, 129:369–385.
- [Ludwig et al., 2002] Ludwig, H.-G., Allard, F., and Hauschildt, P. H. (2002). Numerical simulations of surface convection in a late M-dwarf. *Astronomy and Astrophysics*, 395:99–115.
- [Ludwig et al., 1999] Ludwig, H.-G. et al. (1999). A calibration of the mixing-length for solar-type stars based on hydrodynamical simulations. I. Methodical aspects and results for solar metallicity. *Astronomy and Astrophysics*, 346:111–124.
- [Magic et al., 2013] Magic, Z., Collet, R., Hayek, W., and Asplund, M. (2013). The Stagger-grid: A grid of 3D stellar atmosphere models. II. Horizontal and temporal averaging and spectral line formation. *Astronomy and Astrophysics*, 560:A8.
- [Marchwinski et al., 2015] Marchwinski, R. C., Mahadevan, S., Robertson, P., Ramsey, L., and Harder, J. (2015). Toward Understanding Stellar Radial Velocity Jitter as a Function of Wavelength: The Sun as a Proxy. *The Astrophysical Journal*, 798:63.
- [Mary and Ferrari, 2014] Mary, D. and Ferrari, A. (2014). A non-asymptotic standardization of binomial counts in higher criticism. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 561–565.
- [Masset and Papaloizou, 2003] Masset, F. S. and Papaloizou, J. C. B. (2003). Runaway Migration and the Formation of Hot Jupiters. *The Astrophysical Journal*, 588:494–508.

- [Mayor et al., 2003] Mayor, M. et al. (2003). Setting New Standards with HARPS. *The Messenger*, 114:20–24.
- [Mayor and Queloz, 1995] Mayor, M. and Queloz, D. (1995). A Jupiter-mass companion to a solar type star. *Nature*, 378:355–359.
- [McSweeney, 2006] McSweeney, L. (2006). Comparison of periodogram tests. *Journal of Statistical Computation and Simulation*, 76(4):357–369.
- [Meunier et al., 2010] Meunier, N., Desort, M., and Lagrange, A.-M. (2010). Using the Sun to estimate Earth-like planets detection capabilities . II. Impact of plages. *Astronomy and Astrophysics*, 512:A39.
- [Meunier et al., 2015] Meunier, N. et al. (2015). Using the Sun to estimate Earth-like planet detection capabilities. VI. Simulation of granulation and supergranulation radial velocity and photometric time series. *Astronomy and Astrophysics*, 583:A118.
- [Meunier et al., 2017a] Meunier, N. et al. (2017a). Variability of stellar granulation and convective blueshift with spectral type and magnetic activity. I. K and G main sequence stars. *Astronomy and Astrophysics*, 597:A52.
- [Meunier and Lagrange, 2013] Meunier, N. and Lagrange, A.-M. (2013). Using the Sun to estimate Earth-like planets detection capabilities. IV. Correcting for the convective component. *Astronomy and Astrophysics*, 551:A101.
- [Meunier et al., 2017b] Meunier, N., Lagrange, A.-M., and Borgniet, S. (2017b). A new method of correcting radial velocity time series for inhomogeneous convection. *Astronomy and Astrophysics*, 607:A6.
- [Morbidelli et al., 2005] Morbidelli, A. et al. (2005). Chaotic capture of Jupiter’s Trojan asteroids in the early Solar System. *Nature*, 435:462–465.
- [Mortier et al., 2015] Mortier, A. et al. (2015). BGLS: A Bayesian formalism for the generalised Lomb-Scargle periodogram. *The Astrophysical Journal*, 573:A101.
- [Moscovich et al., 2016] Moscovich, A. et al. (2016). On the exact Berk-Jones statistics and their p-value calculation. *Electronic Journal of Statistics*, 10(2):2329–2354.
- [Moscovich-Eiger and Nadler, 2015] Moscovich-Eiger, A. and Nadler, B. (2015). Fast calculation of boundary crossing probabilities for Poisson processes. *ArXiv e-prints (V.3)*.
- [Muterspaugh et al., 2010] Muterspaugh, M. W. et al. (2010). The phases differential astrometry data archive. v. candidate substellar companions to binary systems. *The Astronomical Journal*, 140(6):1657.
- [Nadler and Kontorovich, 2011] Nadler, B. and Kontorovich, A. (2011). Model selection for sinusoids in noise: Statistical analysis and a new penalty term. *IEEE Transactions on Signal Processing*, 59(4):1333–1345.
- [Nelson et al., 2000] Nelson, R. P., Papaloizou, J. C. B., Masset, F., and Kley, W. (2000). The migration and growth of protoplanets in protostellar discs. *MNRAS*, 318:18–36.
- [Neyman, 1937] Neyman, J. (1937). A smooth test for goodness of fit. *Scandinavian Actuarial Journal*, 1937(3-4):149–199.
- [Nicholls, 1967] Nicholls, D. (1967). Estimation of the spectral density function when testing for a jump in the spectrum. *Australian Journal of Statistics*, 9:103–108.

- [Nicholls, 1969] Nicholls, D. F. (1969). Testing for a jump un co-spectra. *Australian Journal of Statistics*, 11(1):7–13.
- [Nordlund, 1982] Nordlund, A. (1982). Numerical simulations of the solar granulation. I - Basic equations and methods. *Astronomy and Astrophysics*, 107:1–10.
- [Nordlund et al., 2009] Nordlund, Å. et al. (2009). Solar Surface Convection. *Living Reviews in Solar Physics*, 6:2.
- [Nordlund and Galsgaard, 1995] Nordlund, Å. and Galsgaard, K. (1995). A 3d mhd code for parallel computers. *Technical report, Astronomical Observatory, Copenhagen University*.
- [November et al., 1981] November, L. et al. (1981). The detection of mesogranulation on the sun. *The Astrophysical Journal*, 245:L123–L126.
- [Olsen and Bohr, 2010] Olsen, K. and Bohr, J. (2010). Pair-correlation analysis of HD 10180 reveals a possible planetary orbit at about 0.92 AU. *ArXiv e-prints*.
- [Paparoditis and Politis, 1999] Paparoditis, E. and Politis, D. (1999). The local bootstrap for periodogram statistics. *Journal of Time Series Analysis*, 20(2):193–222.
- [Paris, 2013] Paris, S. (2013). *Sparsity-based detection strategies for faint signals in noise : application to astrophysical hyperspectral data*. Theses, Université Nice Sophia Antipolis ; Università degli studi La Sapienza (Rome).
- [Parsen, 1962] Parsen, E. (1962). On estimation of a probability density function and mode. “*The Annals of Mathematical Statistics*”, 33(3):1065–1076.
- [Parzen, 1975] Parzen, E. (1975). *Multiple Time Series: Determining the Order of Approximating Autoregressive Schemes*. Technical report (State University of New York at Buffalo. Statistical Science Division). State University of New York at Buffalo, Department of Computer Science, Statistical Science [Division].
- [Pepe et al., 2010] Pepe, F. et al. (2010). ESPRESSO: the Echelle spectrograph for rocky exoplanets and stable spectroscopic observations. In *SPIE Conference Series*, volume 7735.
- [Pepe et al., 2011] Pepe, F. et al. (2011). The HARPS search for Earth-like planets in the habitable zone. I. Very low-mass planets around HD 20794, HD 85512, and HD 192310. *Astronomy and Astrophysics*, 534:A58.
- [Perryman, 2011] Perryman, M. (2011). *The exoplanet handbook*. Cambridge University Press.
- [Perryman et al., 2014] Perryman, M. et al. (2014). Astrometric Exoplanet Detection with Gaia. *The Astrophysical Journal*, 797:14.
- [Perryman et al., 2001] Perryman, M. A. C. et al. (2001). GAIA: Composition, formation and evolution of the Galaxy. *Astronomy and Astrophysics*, 369:339–363.
- [Petigura et al., 2013] Petigura, E. et al. (2013). Prevalence of earth-size planets orbiting sun-like stars. *Proceedings of the National Academy of Sciences*, 110(48):19273–19278.
- [Plavchan et al., 2015] Plavchan, P. et al. (2015). Radial Velocity Prospects Current and Future: A White Paper Report prepared by the Study Analysis Group 8 for ExoPAG. *ArXiv*.
- [Politis and Romano, 1994] Politis, D. and Romano, J. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89:1303–1313.

- [Priestley, 1981] Priestley, M. (1981). *Spectral Analysis and Time Series*. Academic Press, San Diego.
- [Púciċ et al., 2012] Púciċ, J. et al. (2012). Observability of Spectral Components beyond Nyquist Limit in Nonuniformly Sampled Signals. *ISRN Signal Processing*, 2012:10.
- [Queloz et al., 2001] Queloz, D. et al. (2001). No planet for HD 166435. *Astronomy and Astrophysics*, 379:279–287.
- [Quinn, 1986] Quinn, B. G. (1986). Testing for the presence of sinusoidal components. *Journal of Applied Probability*, 23:201–210.
- [Quinn, 1999] Quinn, B. G. (1999). A fast efficient technique for the estimation of frequency: Interpretation and generalisation. *Biometrika*, 86(1):213–220.
- [Quinn and Fernandes, 1991] Quinn, B. G. and Fernandes, J. M. (1991). A fast efficient technique for the estimation of frequency. *Biometrika*, 78(3):489–497.
- [Quinn and Hannan, 2001] Quinn, B. G. and Hannan, E. J. (2001). *The Estimation and Tracking of Frequency*. Cambridge University Press.
- [Quirrenbach et al., 2014] Quirrenbach, A. et al. (2014). CARMENES instrument overview. In *Ground-based and Airborne Instrumentation for Astronomy V*, volume 9147 of *Proceedings of the SPIE*, page 91471F.
- [Rajpaul et al., 2016] Rajpaul, V. et al. (2016). Ghost in the time series: no planet for Alpha Cen B. *MNRAS*, 456:L6–L10.
- [Ranc et al., 2015] Ranc, C. et al. (2015). MOA-2007-BLG-197: Exploring the brown dwarf desert. *Astronomy and Astrophysics*, 580:A125.
- [Rappaport et al., 2013] Rappaport, S., Sanchis-Ojeda, R., Rogers, L. A., Levine, A., and Winn, J. N. (2013). The Roche Limit for Close-orbiting Planets: Minimum Density, Composition Constraints, and Application to the 4.2 hr Planet KOI 1843.03. *The Astrophysical Journal*, 773:L15.
- [Rayner et al., 2010] Rayner, J. C. W., Thas, O., and Best, D. J. (2010). *Smooth Tests of Goodness of Fit Using R*. John Wiley & Sons, Ltd.
- [Reiners, 2009] Reiners, A. (2009). Activity-induced radial velocity jitter in a flaring M dwarf. *Astronomy and Astrophysics*, 498:853–861.
- [Rieutord and Rincon, 2010] Rieutord, M. and Rincon, F. (2010). The Sun’s Super-granulation. *Living Reviews in Solar Physics*, 7:2.
- [Rieutord et al., 2000] Rieutord, M., Roudier, T., Malherbe, J. M., and Rincon, F. (2000). On mesogranulation, network formation and supergranulation. *Astronomy and Astrophysics*, 357:1063–1072.
- [Rissanen, 1984] Rissanen, J. (1984). Universal coding, information prediction and estimation. *Trans. Inf. Theory*, 30(4):629–636.
- [Roberts et al., 1987] Roberts, D. et al. (1987). Time Series Analysis with Clean - Part One - Derivation of a Spectrum. *The Astrophysical Journal*, 93:968.
- [Robertson et al., 2014] Robertson, P. et al. (2014). Stellar activity masquerading as planets in the habitable zone of the M dwarf Gliese 581. *Science*, 345:440–444.

- [Rouan, 2011] Rouan, D. (2011). Proper motion. In Gargaud, M. et al., editors, *Encyclopedia of Astrobiology*, pages 1345–1347, Berlin, Heidelberg. Springer Berlin Heidelberg.
- [Santos et al., 2004] Santos, N. C. et al. (2004). The HARPS survey for southern extra-solar planets. II. A 14 Earth-masses exoplanet around μ Arae. *Astronomy and Astrophysics*, 426:L19–L23.
- [Scargle, 1982] Scargle, J. (1982). Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, 263:835–853.
- [Schneider, 2005] Schneider, J. (2005). Light-Time Effect and Exoplanets. In Sterken, C., editor, *The Light-Time Effect in Astrophysics: Causes and cures of the O-C diagram*, volume 335 of *Astronomical Society of the Pacific Conference Series*, page 191.
- [Schuster, 1898] Schuster, A. (1898). On the investigation of hidden periodicities. *Journal of Geophysical Research*, 3:13.
- [Schwarzenberg-Czerny, 1998] Schwarzenberg-Czerny, A. (1998). The distribution of empirical periodograms: Lomb–Scargle and PDM spectra. *MNRAS*, 301:831–840.
- [Schwarzenberg-Czerny, 2012] Schwarzenberg-Czerny, A. (2012). Multitaper array processing. In Griffin, E. et al., editors, *Proc. IAU Symp. 285, New Horizons in Time-Domain Astronomy*, page 81. Cambridge Univ. Press,.
- [Seager, 2008] Seager, S. (2008). Exoplanet Transit Spectroscopy and Photometry. *Space Science Reviews*, 135:345–354.
- [Seager, 2013] Seager, S. (2013). Exoplanet habitability. *Science*, 340(6132):577–581.
- [Setiawan et al., 2008] Setiawan, J., Henning, T., Launhardt, R., Müller, A., Weise, P., and Kürster, M. (2008). A young massive planet in a star-disk system. *Nature*, 451:38–41.
- [Shampine, 2008] Shampine, L. (2008). Vectorized adaptive quadrature in matlab. *Journal of Computational and Applied Mathematics*, 211(2):131 – 140.
- [Shao, 2010] Shao, X. (2010). The dependent wild bootstrap. *JASA*, 105:218–235.
- [Shimshoni, 1971] Shimshoni, M. (1971). On Fisher’s test of significance in harmonic analysis. *Geophys J.R. Astronom*, pages 373–377.
- [Shvartzvald et al., 2017] Shvartzvald, Y. et al. (2017). An earth-mass planet in a 1 au orbit around an ultracool dwarf. *The Astrophysical Journal Letters*, 840(1):L3.
- [Siegel, 1980] Siegel, A. (1980). Testing for periodicity in a time series. *Journal of the American Statistical Association*, 75(370):345–348.
- [Smirnov, 1948] Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, 19(2):279–281.
- [Smith, 1985] Smith, R. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72:67–90.
- [Solanki, 2003] Solanki, S. K. (2003). Sunspots: An overview. *The Astronomy and Astrophysics Review*, 11(2):153–286.
- [Soter, 2006] Soter, S. (2006). What Is a Planet? *The Astronomical Journal*, 132:2513–2519.
- [Stein and Nordlund, 2001] Stein, R. and Nordlund, Å. (2001). Solar Oscillations and Convection. II. Excitation of Radial Oscillations. *The Astrophysical Journal*, 546:585–603.

- [Stein, 1967] Stein, R. F. (1967). Generation of Acoustic and Gravity Waves by Turbulence in an Isothermal Stratified Atmosphere. *Solar Physics*, 2:385–432.
- [Stein and Nordlund, 1998] Stein, R. F. and Nordlund, Å. (1998). Simulations of Solar Granulation. I. General Properties. *The Astrophysical Journal*, 499:914–933.
- [Stoica and Moses, 2005] Stoica, P. and Moses, R. (2005). *Spectral analysis of signals*. Prentice Hall.
- [Suleiman et al., 2014] Suleiman, R., Mary, D., and Ferrari, A. (2014). Dimension reduction for hypothesis testing in worst-case scenarios. *IEEE Transactions on Signal Processing*, 62(22):5973–5986.
- [Sulis et al., 2015] Sulis, S., Mary, D., and Bigot, L. (2015). Sur l’utilisation de rapports de périodogrammes pour la détection de raies dans un bruit de densité spectrale partiellement inconnue. In *XXVème colloque GRETSI*, pages 4428–4432.
- [Sulis et al., 2016a] Sulis, S., Mary, D., and Bigot, L. (2016a). Overcoming the stellar noise barrier for the detection of telluric exoplanets: an approach based on hydrodynamical simulations. In Mary, D., Flamary, R., Theys, C., and Aime, C., editors, *EAS Publications Series*, volume 78-79 of *EAS Publications Series*, pages 247–274.
- [Sulis et al., 2016b] Sulis, S., Mary, D., and Bigot, L. (2016b). Using hydrodynamical simulations of stellar atmospheres for periodogram standardization: Application to exoplanet detection. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4428–4432.
- [Sulis et al., 2017a] Sulis, S., Mary, D., and Bigot, L. (2017a). A bootstrap method for sinusoid detection in colored noise and uneven sampling. application to exoplanet detection. In *25th European Signal Processing Conference (EUSIPCO)*, pages 1095–1099.
- [Sulis et al., 2017b] Sulis, S., Mary, D., and Bigot, L. (2017b). A study of periodograms standardized using training datasets and application to exoplanet detection. *IEEE Transactions on Signal Processing*, 65(8):2136–2150.
- [Süveges, 2014] Süveges, M. (2014). Extreme-value modelling for the significance assessment of periodogram peaks. *MNRAS*, 440(3):2099–2114.
- [Süveges et al., 2015] Süveges, M. et al. (2015). A comparative study of four significance measures for periodicity detection in astronomical surveys. *MNRAS*, 450(2):2052–2066.
- [Svensson and Ludwig, 2005] Svensson, F. and Ludwig, H.-G. (2005). Hydrodynamical simulations of convection-related stellar micro-variability. In Favata, F., Hussain, G. A. J., and Battrick, B., editors, *13th Cambridge Workshop on Cool Stars, Stellar Systems and the Sun*, volume 560 of *ESA Special Publication*, page 979.
- [Thibault et al., 2012] Thibault, S. et al. (2012). Spirou@ cfht: spectrograph optical design. In *Proc. of SPIE Vol*, volume 8446, pages 844630–1.
- [Thomas and Weiss, 1992] Thomas, J. and Weiss, N. (1992). *Sunspots: Theory and Observations*, volume 375. Springer Netherlands.
- [Thong et al., 2004] Thong, T. et al. (2004). Lomb-Wech Periodogram for Non-uniform Sampling. *Proc. 26th Annual International Conf. of the IEEE EMBS*.
- [Tingyan, 2010] Tingyan, Y. (2010). Multistep yule-walker estimation of autoregressive models. Master’s thesis, National University of Singapore.

- [Title et al., 1989] Title, A. M. et al. (1989). Statistical properties of solar granulation derived from the SOUP instrument on Spacelab 2. *The Astrophysical Journal*, 336:475–494.
- [Trampedach et al., 2013] Trampedach, R., Asplund, M., Collet, R., Nordlund, Å., and Stein, R. F. (2013). A Grid of Three-dimensional Stellar Atmosphere Models of Solar Metallicity. I. General Properties, Granulation, and Atmospheric Expansion. *The Astrophysical Journal*, 769:18.
- [Truong-Van, 1990] Truong-Van, B. (1990). A new approach to frequency analysis with amplified harmonics. *Journal of the Royal Statistical Society: Series B*, 52(1):203–221.
- [Tsiganis et al., 2005] Tsiganis, K. et al. (2005). Origin of the orbital architecture of the giant planets of the Solar System. *Nature*, 435:459–461.
- [Tukey, 1976] Tukey, J. (1976). The higher criticism. *Course Notes, Statistics 411*.
- [Tuomi et al., 2012] Tuomi, M. et al. (2012). Signals embedded in the radial velocity noise. Periodic variations in the τ Ceti velocities. *Astronomy and Astrophysics*, 551:A79.
- [Tuomi et al., 2014] Tuomi, M. et al. (2014). Bayesian search for low-mass planets around nearby M dwarfs - estimates for occurrence rate based on global detectability statistics. *MNRAS*, 441:1545–1569.
- [Unno et al., 1989] Unno, W. et al. (1989). *Nonradial oscillations of stars*.
- [Van Trees, 2002] Van Trees, H. (2002). *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. John Wiley & Sons, Inc.
- [VanderPlas, 2017] VanderPlas, J. T. (2017). Understanding the Lomb-Scargle Periodogram. *ArXiv e-prints*.
- [Vio et al., 2010] Vio, R., Andreani, P., and Biggs, A. (2010). Unevenly-sampled signals: a general formalism for the Lomb-Scargle periodogram. *Astronomy and Astrophysics*, 519:A85.
- [Vityazev, 1994] Vityazev, V. V. (1994). The Time Interferometer: a New Tool for the Spectral Analysis of Time Series. *Astronomical and Astrophysical Transactions*, 5:177.
- [Vögler et al., 2005] Vögler, A. et al. (2005). Simulations of magneto-convection in the solar photosphere. Equations, methods, and results of the MURaM code. *Astronomy and Astrophysics*, 429:335–351.
- [Vogt, 1987] Vogt, S. S. (1987). The Lick Observatory Hamilton Echelle Spectrometer. *PASP*, 99:1214–1228.
- [Vogt et al., 2015] Vogt, S. S., Burt, J., Meschiari, S., Butler, R. P., Henry, G. W., Wang, S., Holden, B., Gapp, C., Hanson, R., Arriagada, P., Keiser, S., Teske, J., and Laughlin, G. (2015). Six Planets Orbiting HD 219134. *The Astrophysical Journal*, 814:12.
- [Vogt et al., 2010] Vogt, S. S. et al. (2010). The Lick-Carnegie Exoplanet Survey: A 3.1 ME Planet in the Habitable Zone of the Nearby M3V Star Gliese 581. *The Astrophysical Journal*, 723:954–965.
- [Von Sachs, 1993] Von Sachs, R. (1993). Estimating the spectrum of a stochastic process in the presence of a contaminating signal. *Signal Processing, IEEE Transactions on*, 41(1):323.
- [Von Sachs, 1994] Von Sachs, R. (1994). Peak-insensitive non-parametric spectrum estimation. *Journal of Time Series Analysis*, 15(4):429–452.

- [Wakefield, 2013] Wakefield, J. (2013). *Bayesian and Frequentist Regression Methods*. Springer, New York.
- [Walther, 2012] Walther, G. (2012). The Average Likelihood Ratio for Large-scale Multiple Testing and Detecting Sparse Mixtures. *IMS Collections : From Probability to Statistics and Back: High-Dimensional Models and Processes*, 9:317–326.
- [Welch, 1967] Welch, P. D. (1967). The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15:70–73.
- [White, 1999] White, L. (1999). Detection of sinusoids in unknown coloured noise using ratios of ar spectrum estimates. *Proc. Information, Decision and Control*, pages 257–262.
- [Whittle, 1952] Whittle, P. (1952). The simultaneous estimation of a time series harmonic components and covariance structure. *Trabajos de Estadística*, 3(1-2):43–57.
- [Whittle, 1954] Whittle, P. (1954). The statistical analysis of a seiche record. *Sears Foundation Journal of Marine Research*, 13:76–100.
- [Williams and Pesnell, 2011] Williams, P. E. and Pesnell, W. D. (2011). Comparisons of supergranule characteristics during the solar minima of cycles 22/23 and 23/24. *Solar Physics*, 270(1):125–136.
- [Williams and Pesnell, 2014] Williams, P. E. and Pesnell, W. D. (2014). Time-Series Analysis of Supergranule Characteristics at Solar Minimum. *Solar Physics*, 289:1101–1113.
- [Wolszczan and Frail, 1992] Wolszczan, A. and Frail, D. (1992). A planetary system around the millisecond pulsar PSR1257 + 12. *Nature*, 355:145–147.
- [Young, 1994] Young, G. (1994). Bootstrap: More than a stab in the dark? *Statistical Science*, 9(3):382–395.
- [Zechmeister and Kürster, 2009] Zechmeister, M. and Kürster, M. (2009). The generalised Lomb-Scargle periodogram. A new formalism for the floating-mean and Keplerian periodograms. *Astronomy and Astrophysics*, 496:577–584.
- [Zheng, 2012] Zheng, C. (2012). Detection of multiple sinusoids in unknown colored noise using truncated cepstrum thresholding and local signal-to-noise-ratio. *Applied Acoustics*, pages 809–816.
- [Zoubir and Iskander, 2004] Zoubir, A. and Iskander, R. (2004). *Bootstrap Techniques for Signal Processing*. Cambridge University Press.