ÉCOLE POLYTECHNIQUE

PROMOTION X2017

RICHARD Pablo

RESEARCH INTERNSHIP REPORT

Improvement of exoplanets research methods

Unmixing exoplanetary signals from stellar activities in radial velocity measurements using independent component analysis

NON-CONFIDENTIAL

Option PHY592 : Astrophysics and cosmology
Field : Exoplanets research

Referring teacher : Frédéric Daigne
Internship supervisors : Rodrigo Díaz and Alain Lecavelier des Étangs

27/04/2020 – 14/08/2020

Host organization : Universidad Nacional de General San Martín
25 avenue de Mayo, 1650 Buenos Aires, ARGENTINA

Déclaration d'intégrité relative au plagiat

Je soussigné RICHARD Pablo certifie sur l'honneur :

1. Que les résultats décrits dans ce rapport sont l'aboutissement de mon travail.
2. Que je suis l'auteur de ce rapport.
3. Que je n'ai pas utilisé des sources ou résultats tiers sans clairement les citer et les référencer selon les règles bibliographiques préconisées.

Je déclare que ce travail ne peut être suspecté de plagiat.

18/08/2020

Résumé

Grâce aux progrès techniques conséquents qu'a connu l'astronomie depuis la seconde moitié du XX^{e} siècle, la découverte de planètes en dehors du Système solaire est passée du stade de fantasme humain à celui de domaine de recherche scientifique en plein essor. La méthode des vitesses radiales, qui se propose de mesurer par spectroscopie le mouvement d'une étoile induit par sa planète en orbite, est un outil majeur qui a permis de nombreuses découvertes. L'un des enjeux principaux est alors de raffiner la résolution des spectromètres dans le but d'estimer avec précision le décalage spectral de l'étoile dû à l'effet Doppler, afin de détecter les potentielles traces de la présence d'exoplanète. Mais il ne s'agit pas du seul facteur limitant. L'activité stellaire en est un autre qui, désormais, devient le nouveau point de résistance dans la détectection d'exoplanètes plus légères, discrètes. En effet l'estimation de la vitesse radiale en théorie due à la planète, est entachée de nombreuses fluctuations dues aux variations spectrales. Celles-ci sont fortement liées à l'activité stellaire magnétique mais également aux différentes tâches sombres présentes à la surface de l'étoile et tournant avec celle-ci. Il en résulte une perturbation, qui peut s'avérer importante, de l'estimation de la vitesse radiale de l'étoile, par différents différents signaux stellaires inhérents à la rotation propre de l'étoile, mais indépendants de l'orbite planétaire. L'objectif est alors d'expliquer les variations spectrales comme le résultat de l'influence combinée de sources stellaires et du signal dû à la planète. Ce type de problème est connu en mathématiques sous le nom de "séparation aveugle de sources", et différentes méthodes peuvent être employées selon les caractéristiques des données étudiées. L'objectif de ce stage fut d'estimer le potentiel d'application de la méthode d'analyse en composantes indépendantes (ICA) aux données de vitesses radiales. Après avoir implémenté et comparé différentes heuristiques d'ICA sur des cas d'école, la première tentative d'application aux données des deux exoplanètes HD 189733 et HD 12484 s'est avérée être un échec. La cause majeure que j'ai identifiée est le compromis à établir entre la réduction de dimension du signal, favorisant la convergence de l'ICA, et la conservation de sa richesse informationnelle, facteur limitant directement l'espace de recherche et donc la meilleure précision possible atteignable par l'ICA. Un travail d'estimation de la qualité des résultats de l'ICA a permis d'intégrer l'ICA dans un processus stochastique avec un module de post-traitement très sélectif, peu sensible aux défauts de convergence de l'ICA. Combiné à une réduction de dimension *ad hoc*, regroupant les ordres spectraux par paquets, j'ai pu obtenir une amélioration du rapport signal sur bruit pour l'estimation de vitesse radiale induite par la planète de 74 à 109 pour HD 189 et de 13 à 27 pour HD 124.

Mots-clés — détection d'exoplanètes, méthode des vitesses radiales, activité stellaire, traitement du signal, analyse en composantes indépendantes, information mutuelle, optimisation géométrique sur les variétés.

Abstract

Thanks to the technical progress that impacted astronomy in the late 20th century, discovering planets outside the Solar System has shifted from a fantasy status to a growing field of scientific research. Radial-velocity method, that suggests to perform spectroscopy measurements of the star motion to deduce the planet orbit, has revealed to be a very powerful tool to detect new exoplanets. In order to track with precision a spectrum shift, induced by the Doppler effect, and seek in it any planet induced signal, spectrographs have to maintain an extremely high resolution. But this is no longer the main limiting factor. Stellar activity is now becoming an actual barrier as trying to find Earth-like lightweight exoplanets. Indeed, planet induced radial-velocity estimation is corrupted by numerous other independent spectral variations, strongly linked to stellar activity such as magnetic cycles and inherent rotation of the stellar surface containing dark spots. The goal is then to build a model of the measured spectral variations as a combined result of various stellar signals, and the planet induced Doppler shift. This is a signal processing field referred as Blind Source Separation (BSS), where the goal is to describe the observed data as a mixture of some sources without having explicit quantitative information either on the actual sources or the mixing process that lead to the measurements. Depending on the data features, various BSS heuristics can be considered. The goal of this internship was to estimate the potential of Independent Component Analysis (ICA) method when applied to radial-velocity data. A first part of the project was dedicated to implement and compare various ICA heuristics on textbook cases and personal benchmarks. The first attempt on real data of exoplanets HD 189733 and HD 12484 was a failure. The main explication I have found results from the compromise that has to be made between reducing the data dimension, that is necessary for meaningful ICA convergence, and keeping enough signal diversity, otherwise any linear unmixing process is doomed to failure. A special focus on estimating the quality of an ICA result allowed me to build a stochastic ICA pipeline joined with a very selective post-processing module that is sparsely affected by ICA convergence failure. Combined with an *ad hoc* dimension reduction based on spectrum orders grouping, I was able to increase the planet induced radial velocity signal to noise ratio from 74 to 109 for HD 189 and from 13 to 27 for HD 124.

Keywords — exoplanets research, radial velocity method, stellar activity, signal processing, independent component analysis, mutual information, manifold optimization.

Acknowledgements

First, I would like to express my gratitude to my internship supervisors, Rodrigo Díaz and Alain Lecavelier des Étangs, for their trust, benevolence and for their advice and follow-up during this few months of internship.

Alain Lecavelier was my teacher of relativity during the second year at École polytechnique. As I was seeking a research internship in astrophysics in Latin America, Alain helped me in this process. He advised me and gave me a valuable list of contacts, including a strong recommendation to work with Rodrigo Díaz. As my plans of working in Buenos Aires with Rodrigo at the National University of General San Martín were aborted, Alain spontaneously offered me a solution where my internship would be jointly supervised by Rodrigo and himself. I thank Alain for this benevolence and for its guidance during my internship.

As a young father, Rodrigo Díaz managed to allow me some of his precious time. Frequently, after busy working days, he was able to extend his schedule and exchange with me when being at home next to his children. To initiate the project, Rodrigo took the time to prepare especially for me a notebook simulating the data I would have to deal with later on. Conversely, he was always interested in my work when we were interacting. He would systematically and carefully read my notebooks when I wanted his opinion.

It was always a pleasure and a pride to exchange with Rodrigo and Alain, explaining my work, receiving their feedback and their encouragement. Even though both their schedule were very busy and did not allow us to have as many meetings as we wanted, I was able to progress with high autonomy every time after receiving their advice.

I am also grateful to Guillaume Hébrard for its advisory in the exoplanet choice for the application of this project, along with Alain Lecavelier. Their help in finding the corresponding data was very useful.

Finally, I would like to thank Frédéric Daigne who monitored this internship and kept an eye on its smooth running.

Table des matières

Introduction

What is an exoplanet ? Constructed from the Ancient Greek prefix *exo*, standing for 'outside', 'extrasolar', an exoplanet is by definition an extrasolar planet. This designation raises spontaneously the difficulty that will stem from exoplanet detection : they are dark, lightweight and small objects located far away from us. How far ? The closest star to our Solar System is Proxima Centauri, located at 4.25 light years away. It is orbited by Proxima Centauri b, whose existence was confirmed in 2016 and thus ranked it as the closest exoplanet. Within 10 parsecs (32.6 light-years), we have detected less than a hundred exoplanets, distributed over 60 stars approximately. How dark ? A planet is not a direct source of light, its brightness is coming from the lighting of its host star and its image is diluted in the stellar glare. For instance, in the Solar System, the Sun is more than 10^9 times brighter than any other planet lighting. Hence exoplanet detection is, in a certain way, doomed to indirect clues analysis onto its parent star. Nevertheless, some direct methods have recently been employed to image close exoplanets and will probably know a significant improvement in the following decade thanks to technological refinement. For instance, the interferometric instrument GRAVITY of the European Southern Observatory's Very Large Telescope (VLT) in Chile permitted recently the first direct detection of an exoplanet, HR 8799e, 129 light-years away [1].

The key to find an exoplanet is to focus on its parent star and seek for some planet stamp. They are two main methods to search exoplanets : radial-velocity (RV) measurements and transit light curves [2]. The radial-velocity method is based on the movement of the parent star : as the planet is orbiting, the star is also moving around the system center of mass. A projection of this stellar movement can be measured by looking to its spectra, thanks to the Doppler effect. In the end, an estimation of the planet orbital period P_{pl} and its minimum mass $M_{pl}sin(i)$ can be carried out. If the orbital inclination is such that the planet, as seen from Earth, passes in front of the stellar disk (*i.e.* $sin(i) \sim 1$), a slight dimming of the star, called 'transit', is expected. Such transit events carry many information about the star-planet system. Both the actual mass and bulk density of the planet can be derived. Furthermore, it makes possible to study the atmosphere of the transiting planet, both in spectrum and polarization [3]. This is a first step towards comparative exoplanetology. The exciting challenge hidden behind this is the discovering of Earth-like exoplanets, with physical conditions suitable for the complex chemistry of life to develop. Historically, Doppler spectroscopy was the first technique used to reveal the existence of extrasolar planetary systems hosted by solar-type stars [4]. Until 2012, radial-velocity surveys led to the detection of a rich population of exoplanets and was by far the most productive technique used by planet hunters. The Kepler spacecraft finally overtook it in number, using the transit method. Nevertheless, Doppler spectroscopy remains strongly powerful, and are still being improved. As I am writing this report, no less than 4 326 planets distributed among 3 196 planetary systems have been found and confirmed by scientific community.

However, if the 'Holy Grail' quest of Earth-like exoplanets appears as a fantasy, it remains a real scientific challenge. Whether we consider radial-velocities or transit light curves, the difficulty encountered to detect an exoplanet is directly linked to the signal-to-noise ratio (SNR) of the planet induced signal. The planet stamp cannot stand out if the mass ratio M_{pl}/M_s is to slow. Detecting low mass exoplanets has always been a significant challenge. Starting by 51 Pegasi b, in the family of so-called hot Jupiters. It was the first detected exoplanet hosted by a solar-type star, found with the new ELODIE spectrograph at Haute-Provence Observatory in April 1994 as part of a systematic survey of 142 solar-type stars initiated by Mayor and Queloz [5]. In 2004, first Neptune-mass exoplanets have been detected and in 2012, Earth-mass have been reached [4]. If the SNR in RV method was first hampered by instrumental precision, it dropped to 15 $\mathrm{m \cdot s^{-1}}$ in the 1980s and is today reaching the 1 $\mathrm{m \cdot s^{-1}}$. This is no longer a hurdle and SNR is mainly limited by stellar influence on RV estimations. This is why many algorithms to compute RVs out of spectrums measurements have been developed these past 20 years. They can be grouped depending on their complexity and choices for the model of the reference spectra, number of model parameters, and statistics. They can range from simple crosscorrelation with binary masks, to least-squares fit with coadded templates [6] and modelling spectral line spread functions or bayesian inference based on Gaussian processes fitting [7].

The aim of my work is to propose a new RV estimator, without any aid or fitting to the theoretical RV curve. It would be plugged at the end of the RV processing pipeline, taking for input the RV estimations carried out by the Cross-correlation Function technique. Such a 'blind' method can therefore be applied to new unknown systems where the presence of one are various exoplanet is still to be detected.

Chapitre 1

RV in details

This internship was dedicated to improve RV measures processing, in the context of exoplanet detection. In this section, we will investigate in details how RV are estimated out of spectrum records and how stellar activity might disturb this process.

1.1 RV measure and planet signal

1.1.1 Doppler effect

As the distance that separates us to the host star is changing, the Earth measured frequency of the star emitted light is also changing, with respect to the Doppler effect. When the star is moving towards us, its frequency is augmented or equivalently, its wavelength λ is reduced. This is known as the 'blue shift' effect, as a reduction of the direct observable effect on the visible spectrum and is roughly vulgarized by a blue star, as portrayed in figure 1.1. Conversely, when the star is moving away from us, its spectrum is 'red shifted'. Hence, the radial velocity of the star (with respect to Earth) is encoded in its spectrum shift. For two given spectrums of the star, at instants t_1 and t_2, the shift $\Delta\lambda := \lambda(t_2) - \lambda(t_1)$ between the two measured wavelengths of the same monochromatic stellar source is directly related to the stellar radial velocity variation $\Delta RV := RV(t_2) - RV(t_1)$. Making the non-relativistic approximation and denoting by $\bar{\lambda} := (\lambda_1 + \lambda_2)/2$ the averaged wavelength, gives the very simple equation :

$$\Delta RV = c\frac{\Delta\lambda}{\bar{\lambda}} \tag{1.1}$$

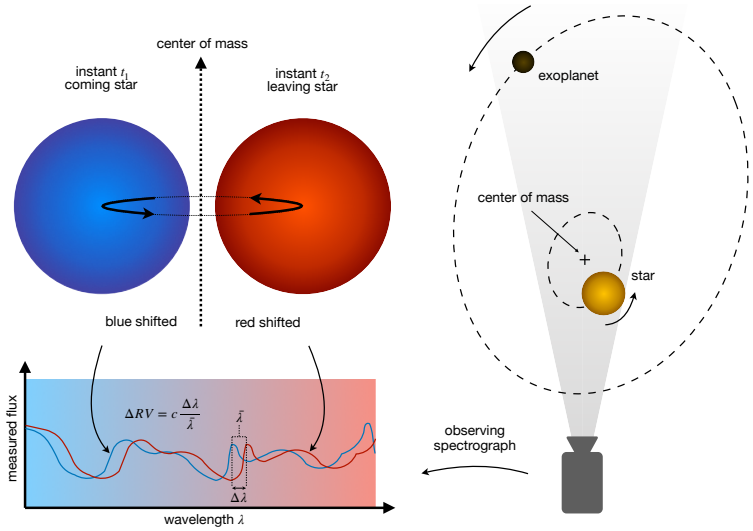


FIGURE 1.1 – Illustration of the Doppler effect induced by an exoplanet orbiting around its star. The spectrum is either blue shifted or red shifted whether the star is moving forward or away from us.

1.1.2 Experimental performance of échelle grating and CCD

As explained above, radial velocities are estimated out from spectrum shiftings. The computation accuracy is thus preconditioned by the spectrograph precision : both vertical (in light flux per wavelength) and horizontal (wavelength calibration). As the ratio M_{pl}/M_s is usually very low, the planet induced radial velocities RV_{pl} are challenging to detect. For instance Jupiter has a RV effect of $13\,\mathrm{m\cdot s^{-1}}$ over the Sun over a 12 years period and Earth's influence is no more than $9\,\mathrm{cm\cdot s^{-1}}$. The main high-precision échelle planet-finding spectrograph used today is the High Accuracy Radial Velocity Planet Searcher (HARPS), installed in 2002 on the ESO's 3.6 m telescope at La Silla Observatory in Chile and provides state-of-the-art [1] of stellar RV

1. Another spectrograph, Echelle Spectrograph for Rocky Exoplanet and Stable Spectroscopic Observations (ESPRESSO), was mounted, highly inspired by HARPS, on the European Southern Observatory's VLT. It was designed to reach the $1\,\mathrm{cm\cdot s^{-1}}$ precision

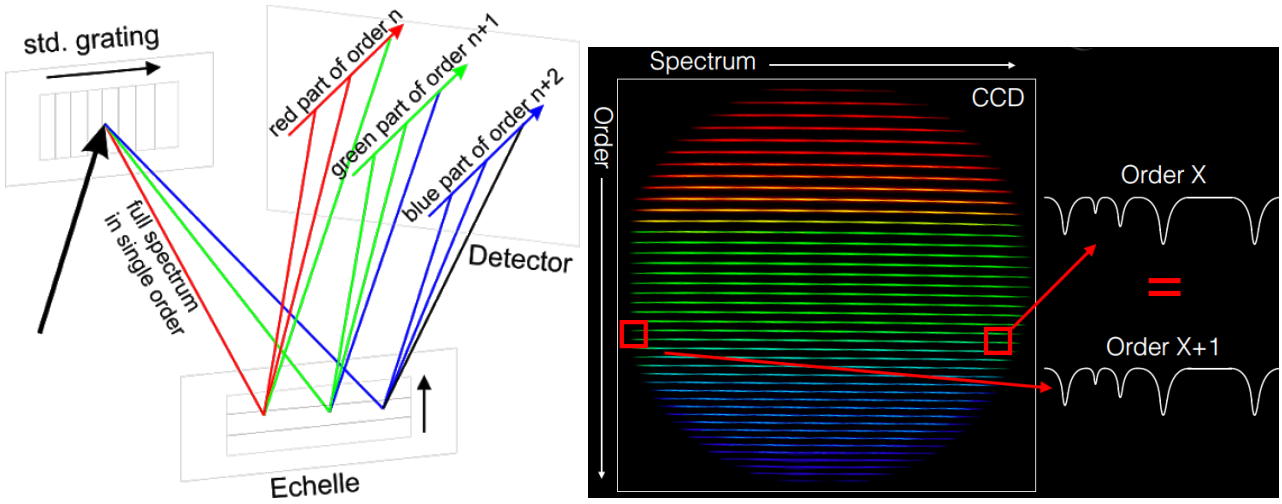
measurements with a precision down to $1\,\mathrm{m}\cdot\mathrm{s}^{-1}$. In this internship, we investigated RV estimations of HD 189 and HD 124 made out of spectrums recorded by the SOPHIE high-resolution échelle spectrograph installed on the 1.93 m reflector telescope at the Haute-Provence Observatory located in south-eastern France, reaching the same order of precision.

When applying spectral decomposition to the star light, the luminous flux is diffracted in order to project each spectral line on a specific pixel. Hence, each pixel detector receives a very tiny part of the initial flux. The detector is being put under a faint light regime where quantization appears and generates a photon noise. The flux per pixel, seen as a number N_{pix} of received photons for a given exposure time τ, is a random variable following a Poisson distribution :

$$\mathbb{P}(N_{\mathrm{pix}} = n) = \frac{\tau\mu_{\mathrm{pix}}}{n!} e^{-\tau\mu_{\mathrm{pix}}} \tag{1.2}$$

where μ_{pix} is a certain rate depending on the enlightment intensity of the pixel. To deal with this faint light regime, charge-coupled device (CCD) captors are usually used because of their high quantum efficiency an can be easily assembled in 2D arrays. The SNR is then limited by the exposure time, in proportion to $1/\sqrt{\tau}$. A deeper analysis of photon noise impact over radial velocity measurements was made by Bouchy et al. in [8] who set up a methodology to compute the fundamental limit it brings.

Equation (1.1) shows us that RV precision is determined by the resolution of the spectrograph $R := \lambda/\delta\lambda$. The best spectral resolution $R := \lambda/\delta\lambda$ are of order 100 000 and are reached using échelle spectrographs. Their principle is illustrated in figure 1.2. The CCD images containing the raw SOPHIE spectral information have 2154×4102 pixels [2]. Asking a RV precision of $1\,\mathrm{m}\cdot\mathrm{s}^{-1}$ requires to track a spectrum shift $\Delta\lambda$ of the order 10^{-5} Å if working around visible spectrum band. This represents a screen-shift of about 1/1000 of a CCD pixel. These spectrographs need to be extremely stabilized in pression and temperature. In general, a fiber joins the focal point of the telescope to the spectrograph that is monitored in a specific shield. This allow to avoid any air refraction and slit illumination variation effect. In addition to this, calibration signals are also thrown into the spectrograph to correct from other drifts. Finally, additive drifts in RV such as Earth movement are removed [9].



(a) échelle grating principle

(b) Decomposed spectrum view over the HARPS CCD screen

.

FIGURE 1.2 – Illustration of the échelle grating principle and results. The incident beam is diffracted first by the standard grating, then by the échelle grating. The full spectrum is split into various orders, finally caught by the CCD matrix.
Credit (a) — 'Echelle Principle' by Boris Považay (Cardiff University) - Own work. Licensed under CC BY-SA 2.5 *via* Wikimedia Commons.
Credit (b) — X. Dumusque presentation 'Radial Velocity Surveys' at 2015 Sagan Exoplanet Summer Workshop, https ://vimeopro.com/vcubeusa/caltech-2015/page/3 [10]

However, since it saw its first light in 2016, many technical issues impeded its using until now.
2. More information about SOPHIE can be found here : http ://www.obs-hp.fr/guide/sophie/data_products.shtml

label	value	description
seq	406300	Sequence number
objname	HD189733	Designation of the object
date	2007-07-19	Observation date
slen	1	Length of the series
mask	G2	Name of the mask
ccf_offline	0	Offline CCF
rv	-2.147 km/s	Radial velocity
fwhm	8.720 km/s	FWHM of the correlation peak
maxcpp	1811	Max. number of counts per pixel
contrast	41.09 %	Contrast of the CCF
lines	4736	Nbr. of lines used for the CCF

TABLE 1.1 – Metadata of the corresponding CCF

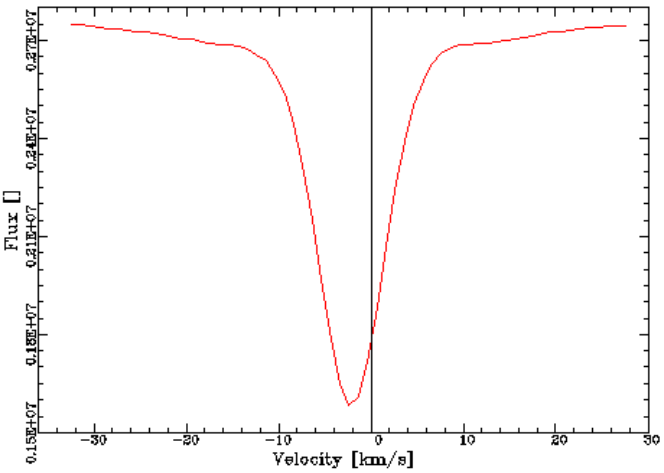


FIGURE 1.3 – Exemple of a CCF available on the database. This is HD 189 observed by SOPHIE the 07/19/2007. ***Credit*** — The SOPHIE archive database

1.1.3 Cross-correlation Function

We have seen previously that RV can be computed by estimating the shift $\Delta\lambda$ of the spectrum with respect to some reference, and that échelle spectrographs offer enough spectral resolution to measure this shift. This estimation process is nevertheless an important part, that will play no mean stake in this project. There are many ways to deal with it. The simplest and traditional one that is often used is the Cross-correlation Function (CCF) method.

When knowing the spectrum profile type of the studied star, one can build a binary window-shaped mask $M(\lambda)$ non-vanishing only over the stellar absorption lines. The observed spectrum of the star, at a given time, $I(\lambda)$ is then cross-correlated with the mask, to give the CCF :

$$\mathrm{CCF}(\delta\lambda) := (I \star M)[\delta\lambda] := \int M(\lambda + \delta\lambda) I(\lambda) \mathrm{d}\lambda \tag{1.3}$$

An exemple [3] of a CCF used to estimate HD 189 radial velocity is given in figure 1.1.3, and the corresponding parameters are reported in table 1.1. Many information can be carried out from a CCF curve. The shift $\Delta\lambda$ is estimated by performing a Gaussian fit onto this curve, and then is derived the RV estimation. The Full Width at Half Maximum (FWHM) combined with the contrast C of the CCF and the SNR of the measured spectrum can be used to derive a proxy for RV precision as follow :

$$\sigma_{RV} \sim \frac{\sqrt{FWHM}}{C.SNR} \tag{1.4}$$

As the total equivalent width must be conserved, C and $FWHM$ are linked : $C \sim 1/FWHM$. From this formula we can see that the broadening of spectral lines, caused either by some stellar activity or any instrumental spectral resolution lack, can be a strong limit for radial velocity precision : any perturbation will simultaneously increase the FWHM and therefore degrade the RV precision with $FWHM^{3/2}$ [11].

It is usually preferred to split the CCF prediction over the various spectrum orders given by the échelle spectrograph rather than analyzing the full spectrum at once. For an order k, this is done by taking the cropped mask M_k onto the specific order segment $[\lambda_k^-, \lambda_k^+]$:

$$\overline{RV}_k := \mathbb{E}\big(\mathrm{GaussianFit}\big[I \star M_k\big]\big) \tag{1.5}$$

Then some *ad hoc* filtering processes can be applied according to specific information collected about the star, like knowing which orders are corrupted by its activity and which one are not. Our goal is precisely to take action here in the pipeline. We propose in this work a new way to post process these multi orders CCF based RV predictions using ICA.

3. Data for HD 189 are available here : http ://atlas.obs-hp.fr/sophie/sophie.cgi ?n=sophiescc&a=htab&ob=ra,seq&c=o&o=HD189733

1.2 Stellar activity

Even if experimental tools have reached an extremely high accuracy level thanks to échelle grating and can be corrected for the various drifts [9], the technical performance is not the only barrier in RV estimation.

1.2.1 Sources of noise and RV effect

Granulation

The star has a convective zone in which the hot gas rises up until it reaches the photosphere, before it cools down and goes backwards toward the centre. Hence, in average, the gas that goes towards the observer is the one rising up to the photosphere, hot and bright, while the one that goes away from observer is for the same reason cooler, so fainter. This causes an overall blueshift of the surface. This effect is usually neglected because of its high granularity, that has few impact on the overall RV estimation.

Spots

Stellar magnetic field can locally block its convective flux. The corresponding regions of the photosphere get cooler than their neighbourhood and thus appear darker : this is called a starspot. Because the magnetic field lines are dragged by the stellar inherent rotation, these spots are rotating with the photosphere and thus cause a variating dim of the measured flux around its low-temperature corresponding wavelength as shown in figure 1.4. This is the source of noise in RV estimation. It is strongly correlated to the stellar inherent rotation period, that may dazzle any planet signal that could be found on the same frequency. Hence hunting starspots [12] and modelling their impact [13] has became an import field of research. An open-source Python package, PSOAP, has been implemented to simulate stellar spectrum corrupted by starspots [14] and infer simultaneously exoplanet orbit and stellar spectra using Gaussian processes [7].

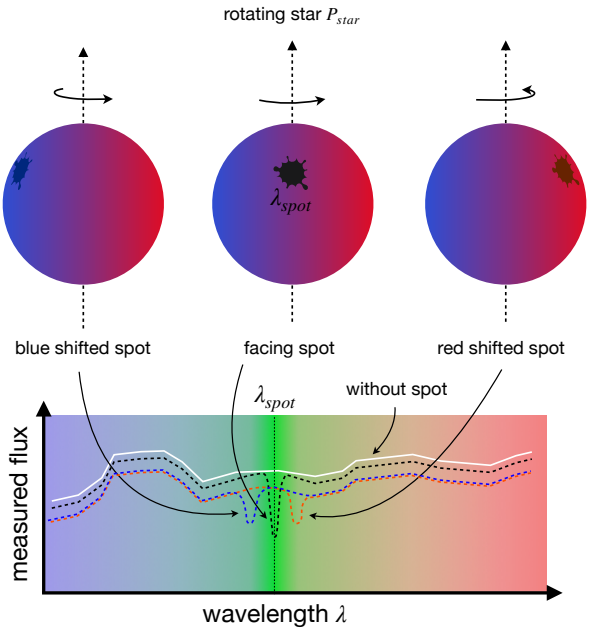


FIGURE 1.4 – Impact of a moving starspot on the spectrum. The flux dim caused by the spot is alternately blue shifted (left figure, blue spectrum), not shifted (middle figure, dark spectrum), red shifted (left figure, red spectrum) are not visible (white spectrum).

1.2.2 Gaussian processes and correction attempt

Gaussian processes are very powerful to model RV perturbations induced by these starspots [2]. They are used to generate time series, out of few parameters that control their covariance function. The choice of these parameters allow to design a certain structure that will be recovered when drawing a random sample out of it. A popular model for the covariance function is given by :

$$k(t_i, t_j) = \sigma_i^2 \delta_{ij} + A^2 \exp\left[-\frac{(t_i - t_j)^2}{2\tau^2} - \frac{2}{\eta^2}\sin^2\left(\frac{\pi(t_i - t_j}{\mathcal{P}}\right)\right] \tag{1.6}$$

where the first term $\sigma_i^2 \delta_{ij}$ accounts for any uncorrelated perturbation such as the instrument noise, and the second term could be thought as a starspot induced signal. Its intensity is monitored by A, and its typical lifetime by τ. During its lifetime, the spot will introduce a noise, strongly correlated with the stellar photosphere rotation of period $\mathcal{P}$.

The objective behind describing the stellar perturbation by Gaussian processes is to be able to infer it. From this simple model, a likely-hood function can be built and then the stellar perturbation can be infered using Markov Chain Monte Carlo (MCMC) algorithms. Once inferred it can be removed to have only the stellar RV.

Chapitre 2

The Independent Component Analysis

2.1 General principle and link with RV perturbations

2.1.1 Facing a BSS problem

The previous part showed how some perturbations such as the stellar activity can disturb the RV estimation. The RV induced planet signal is not the only physical effect measured on the spectrum variation : some orders of the spectrum can be affected by another type of signals. Trying to recover a source signal out of some measurements is a signal processing field known as Blind Source Separation (BSS). The goal is to describe the observed data as a mixture of some sources without having explicit quantitative information either on the actual sources or the mixing process that lead to the measurements. Depending on the qualitative assumptions made about the nature of the mixing process or the criteria that must follow the sources, various BSS heuristics can be considered. Among some of them, one might cite Principal Component Analysis (PCA), Independent Component Analysis (ICA), Dependent Component Analysis (DCA), Wold's decomposition, Non-negative matrix factorization.

2.1.2 ICA as a potential solution

We investigate in this project the potential of ICA to recover the planet signal out of its mixture with stellar activity. The goal of ICA is to solve BSS problems which arise from a linear mixture. A very good tutorial on ICA is given by J. Shlens in [15]. In the following, we will introduce briefly the principle of ICA before presenting and comparing various ICA computation heuristics.

The cocktail party problem, where the objective is to recover, out of two mikes, a specific voice mixed up with a background music, is a typical ICA problem that illustrates well its objective. In such a problem, the mikes are constituting a set of $p = 2$ measure channels $X = \big(X_1(t), \dots, X_p(t)\big)$, supposed to carry out a linear mixture of $l = 2$ independent sources $S = \big(S_1(t), \dots, S_l(t)\big)$, here the voice and the background music. Such a relation can easily be rephrased as :

$$X(t) = AS(t) \tag{2.1}$$

where A is a $l \times p$ real valued matrix, constant in time, accounting for the way signals of S are combined during the measuring process that raised X. The major concern of ICA is to find both an estimation for $\hat{A}$ and $\hat{S}$ that satisfies jointly the equality constraint $X = \hat{A}\hat{S}$ and the mutual independence assumption of sources $\hat{S}_1, \dots, \hat{S}_l$ in $\hat{S}$.

Why are we motivated by ICA to process RV data ? Is it justified to apply such a method in this specific problem ? As I have been asked to seek the potential of ICA processing to RV data, these two questions have naturally arisen to myself, and I think the goal of my internship was more to begin the discussion of these questions than to achieve a successful data processing pipeline, not knowing why it is meant to solve this problem.

First of all, the nature of our problem is quite suitable with ICA principle : the data we are given to analyze is a multivariate time series : splitting the CCF computation over several fixed orders of the full spectrum gives actually a set of measure channels. The RV estimation $X_k := \mathrm{CCF}_k\big(RV_{pl}, \chi\big)$ for order k is influenced by the planet signal RV_{pl}, mixed with several stellar activities signals χ, having each one a specific effect per order. Without any stellar activity $\chi = 0$, this estimation is a random variable of expected value RV_{pl} (systematic errors are supposed to be removed by calibration), with a fluctuation term that we simply model by a gaussian noise $\epsilon_k \sim \mathcal{N}(0, \sigma_{RV_k})$:

$$\mathrm{CCF}_k\big(RV_{pl}, \chi = 0\big) = RV_{pl} + \epsilon_k \tag{2.2}$$

In our case, spectrum orders estimations play the 'mikes' role, as a mixture of planet induced signal RV_{pl} and various stellar activity signals $\chi_1, \dots, \chi_{l-1}$, supposed to be mutually independent. To emphasize this mixture, let us develop the k-th order RV estimation X_k with respect to stellar activity χ :

$$X_k := \mathrm{CCF}_k\big(RV_{pl}, \chi\big) = \mathrm{CCF}_k\big(RV_{pl}, \chi = 0\big) + \nabla_{\chi=0}\mathrm{CCF}_k\big(RV_{pl}\big).\chi + O\big(\|\chi\|^2\big)$$

$$X_k = RV_{pl} + \epsilon_k + \nabla_{\chi=0}\mathrm{CCF}_k\big(RV_{pl}\big).\chi + O\big(\|\chi\|^2\big) \tag{2.3}$$

As we want a linear dependence between X_k and RV_{pl}, we have to approximate the gradient term in equation 2.3 :

$$\nabla_{\chi=0}\mathrm{CCF}_k\big(RV_{pl}\big) = \nabla_{\chi=0}\mathrm{CCF}_k\big(RV_{pl} = 0\big) + O\big(RV_{pl}\big) \tag{2.4}$$

Which leads us to the following linearization, written in a matrix form :

$$X = \overbrace{\begin{pmatrix} 1 & \nabla^T_{\chi=0}\mathrm{CCF}_1\big(RV_{pl}=0\big) \\ \vdots & \vdots \\ 1 & \nabla^T_{\chi=0}\mathrm{CCF}_p\big(RV_{pl}=0\big) \end{pmatrix}}^{A} \overbrace{\begin{pmatrix} RV_{pl} \\ \chi_1 \\ \vdots \\ \chi_{l-1} \end{pmatrix}}^{S} + \epsilon + O\big(RV_{pl}.\|\chi\|\big) + O\big(\|\chi\|^2\big) \tag{2.5}$$

To fit ICA requirements, linearization 2.5 raises 3 points :

- Non-linearity : are the terms hidden in $O\big(RV_{pl}.\|\chi\|\big)$ and $O\big(\|\chi\|^2\big)$ small enough not to disturb ICA ?
- Noise : is ϵ, accounting for measurement induced errors, small enough ?
- Model : is stellar activity χ reducible in few independent time series χ_k ? Are they significant enough to be caught by ICA ?

Since the first point is hard to assess in practice, we made the choice to go for the simplest in this project by giving a try to ICA, under the linear assumption. If it fails, it could possibly be relevant to analyze deeper this point and design a more suitable heuristic, such as non-linear ICA. But couldn't we go for even more simple ? Is it really relevant to try to carry out the full linear decomposition of RV_{pl} and χ from an ICA processing ? Indeed, one could try to recover RV_{pl} out of X by a simple averaging over orders. If all the columns of the mixing matrix A in equation 2.5 sum is close to 0, except the first one, it could be very efficient. The first effect would be to delete the stellar activity influence, but it would also divide the noise ϵ by a factor $\sqrt{p}$ where p is the number of channels used. A very important question is then the following one :

Is there at least one stellar signal χ_k having an influence over RV estimations that cannot be removed by a simple averaging operation over orders ?

Or equivalently, in a quantitative formulation, to what extent the stellar activity respect the following condition :

$$\max_{1\leq k\leq l-1} \Big(\sum_{i=1}^{p} A_{ik}\Big)^2 \mathrm{Var}(\chi_k) \ll p^2\,\mathrm{Var}(RV_{pl}) \tag{2.6}$$

If condition 2.6 is respected, there is no need to call ICA, a simple order averaging will be way more simple and robust. To try to answer this question, we can imagine a stellar signal χ_k as the RV perturbation signal induced by a moving spot on the stellar surface as illustrated in figure 1.4, and wonder if its global effect over RV estimation $\sum_{i=1}^{p} A_{ik}$ could be vanishing. Whether it disturbs one or several orders $k_1, \dots, k_r$, its RV perturbations $\partial_{\chi_k}\mathrm{CCF}_{k_1}\big[RV_{pl}=0, \chi=0\big].\chi_k(t), \dots, \partial_{\chi_k}\mathrm{CCF}_{k_r}\big[RV_{pl}=0, \chi=0\big].\chi_k(t)$ should be, at any time t, of same sign. And, because the time series $\big(\chi_k(t)\big)_t$ is up to a ± 1 factor, we can even impose all coefficients $(A_{ik})_i$ of k-th column of matrix A to be either equal are greater than 0. However, this would no longer be the true if we encode in χ_k the combined RV influence of two spots, of the same stellar latitude, but having an opposite phase : when one of them is visible, the other one is hidden on the other side of the star. To thwart this phenomenon, a possible approach would be to use a BSS method that would seek A, S such that $X = AS$ and A is a non-negative matrix, *i.e.* containing only positive coefficients. But as long as my focus on ICA gave me enough ideas to explore, I was not able to implement and give a try to this other research possibility.

2.2 Implementation of ICA

2.2.1 From time series to random variables

In which context can we apply ICA ? This is a very tricky point, usually not questioned when ICA is performed over time series. ICA is designed to operate over mathematical objects where basic notions like average, variance/power, correlation and most above all independence should be defined and make sense.

The powerful conception of random variables, inherent to the probabilistic paradigm is the ideal tool to define these notions and derive an ICA heuristic. It is directly and naturally followed by the statistical paradigm, in which we interact with the random variables only through a list of independent observations generated by the same distribution. An estimation of the latter can be carried out and the problem is thus reduced to the previous probabilistic point of view. A broader context of application is time series, or even broader : stochastic processes. Two major points distinguish times series to simple statistical analysis : stationarity and ergodicity. Time series are nothing more than a sequence of random variables, each one associated to the result obtained at a specific instant of the series' timeline. A stationary stochastic process, or time series, assumes that all these observations are carried out from the same probability distribution. This is the famous 'identically distributed' hypothesis made when working with *i.i.d.* random variables. The ergodicity assumption is a powerful idea that puts on an equal footing time averaging and probabilistic expectation. This is a natural property given when working on 'independent and identically distributed' variables but false in general when working with time series having any time-lagged correlation term.

In the following, I will present ICA implementation for a statistical context, or equivalently, for stationary and ergodic time series. As mentioned in the first part, Gaussian processes could be used to model stellar activity χ influence over RV data. This suggests explicitly that RV time series have strong time-lagged correlations that violates the ergodicity assumption. Even though I am aware of this major limit, I decided to give a try to my ICA routines.

2.2.2 Pre-processing

Before running into the ICA, RV data has to undergo a preprocessing step. To keep it general, let us name X the measured data. The timeline is made out of n instants $t_1, \dots, t_n$ and for each instant, the measure have p channels. The data can be stored in a $p \times n$ matrix :

$$X = \overbrace{\left\{\begin{array}{ccc} X_1(t_1) & \dots & X_1(t_n) \\ \vdots & \ddots & \vdots \\ X_p(t_1) & \dots & X_p(t_n) \end{array}\right.}^{\text{time axis : } n \text{ instants}} \left.\vphantom{\begin{array}{ccc} X_1(t_1) & \dots & X_1(t_n) \\ \vdots & \ddots & \vdots \\ X_p(t_1) & \dots & X_p(t_n) \end{array}}\right\}_{\text{spectrum axis : } p \text{ orders}} \tag{2.7}$$

Centering

The very first step consists in centering the data around its expected value :

$$X - \mathbb{E}(X) \hookrightarrow X \tag{2.8}$$

Thanks to the ergodicity assumption, the expected value can be estimated as follow :

$$\mathbb{E}(X) \underset{\text{erg.}}{\simeq} \bar{X} := \frac{1}{n} \sum_{j=1}^{n} X(t_j) \tag{2.9}$$

Whitening

Now that the first order has been removed, let us process the second order. Remembering that our goal is to find some mutual independent sources $S = \big(S_1(t), \dots, S_l(t)\big)$ and a matrix A such that $X = AS$, we can see that applying any linear invertible operation $W \in GL_p(\mathbb{R})$ over X will just shift the mixing matrix estimation from A to WA but will not affect sources estimation. Since uncorrelatedness is a necessary condition to independence, it is natural to pre-process X into a whitened version $X_w := WX$ that is uncorrelated. The choice for W can be made using the Principal Component Analysis (PCA), that is nothing more than a Singular Value Decomposition (SVD) of the data matrix : $X = U\Sigma V^T$. Both U and V are asked to be orthogonal matrices and Σ a diagonal matrix. Such a decomposition is easily obtained by looking at the covariance matrix :

$$C := n\big[\operatorname{Cov}(X_i, X_j)\big]_{i,j} \underset{\text{erg.}}{\simeq} XX^T = U\Sigma^2 U^T \tag{2.10}$$

From the real valued and symmetric matrix C orthogonal diagonalization, it is possible [1] to recover both Σ and U up to permutation between signals and a ± 1 factor for each one of them. Then choosing $W = \Sigma^{-1}U^T$ allows to whiten the data :

$$\left[n\,\mathrm{Cov}(X_{w,i}, X_{w,j})\right]_{i,j} \underset{\text{erg.}}{\simeq} X_w X_w^T := WXX^TW^T = \Sigma^{-1}U^TU\Sigma^2U^TU\Sigma^{-1} = I_p \tag{2.11}$$

This SVD decomposition is known as Principal Component Analysis routine because the initial data X has been split into new signals. Writing $X = U\Sigma X_w$ can be rephrased as follow :

$$X = U\sum_{k=1}^{p} \lambda_k X_{w,k} \tag{2.12}$$

where $\Sigma = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$. Equation (2.12) shows us that, up to an orthogonal changement of basis U, the data is composed of p mutually uncorrelated signals $X_{w,1}, \ldots, X_{w,p}$ with respective variances $\lambda_1^2, \ldots, \lambda_p^2$. If we assume that these eigenvalues have been sorted by decreasing amplitude : $\lambda_1^2 > \cdots > \lambda_p^2$, the principal components of X have been identified, up to a ± 1 factor for each component :

rank	**variance**	**normalized signal**
1	λ_1^2	$\pm X_{w,1}(t)$
2	λ_2^2	$\pm X_{w,2}(t)$
$\vdots$	$\vdots$	$\vdots$
p	λ_p^2	$\pm X_{w,p}(t)$

TABLE 2.1 – PCA results overview

Truncating

The previous part brought us many information on the nature of data X. In particular, having the variances distribution $\lambda_1^2, \ldots, \lambda_p^2$ gives us an idea on how is distributed the power of the signal. If we have an estimation of the noise level σ_ϵ that is added by the instrument during the measure : $X = AS + \epsilon$, we can deduce that all signals $X_{w,k}$ derived by the PCA such that $\lambda_k^2 \leq \sigma_\epsilon^2$ are entirely corrupted by the additive noise and therefore should not be considered. Throwing these signals into the ICA pipeline would just impede or even skew its convergence. A simple trick just consists in truncating the PCA and keep only signals with high enough variance.

2.2.3 Independence matter

How to assess independence out of few samples of a multivariate random variable ? This is a tricky question that has various answer possibilities. The independence of some sources $\hat{S} = (\hat{S}_1, \ldots, \hat{S}_p)$ is given by the following criteria over probability distributions :

$$\mathbb{P}_{\hat{S}}(s) = \prod_{k=1}^{p} \mathbb{P}_{\hat{S}_k}(s_k) \tag{2.13}$$

Various independence estimations can be derived from this definition. A thorough analysis, including geometrical representations, has been made by J.F. Cardoso in 2004 [16]. They might differ in sensitivity and in calculus cost, that are two points that will condition ICA performance and thus motive this investigation.

Mutual information

Mutual information is a natural function stemming from information theory to judge how close a distribution is to statistical independence. It measures the departure of two or more variables from statistical independence.

$$I(\hat{S}) = \int \mathbb{P}_{\hat{S}}(s) \log_2 \left(\frac{\mathbb{P}_{\hat{S}}(s)}{\prod_{k=1}^{p} \mathbb{P}_{\hat{S}_k}(s_k)} \right) \mathrm{d}s \tag{2.14}$$

This measure can be thought as the Kullback-Leibler divergence (KLD) from the distribution $\mathbb{P}_{\hat{S}}$ to the product of its marginal distributions [16]. It is always positive and vanishes if and only if the independence criteria (2.13) is respected. Thus, it could be a cost function to minimize in ICA.

1. We assume here they are no multiplicity among C eigenvalues. In practice this is always the case when measuring noisy data. Otherwise, any orthogonal transformation is allowed for each degenerated subspace, bringing another source of ambiguity.

Shannon's 'differential entropy'

Defined as :

$$H(\hat{S}) = -\int \mathbb{P}_{\hat{S}}(s) \log_2 \mathbb{P}_{\hat{S}}(s) \mathrm{d}s \tag{2.15}$$

the Shannon's 'differential entropy', or simply entropy, is linked with mutual information :

$$I(\hat{S}) = \sum_{k=1}^{p} H(\hat{S}_k) - H(\hat{S}) \tag{2.16}$$

It can be used to compute efficiently the mutual information [15]. Indeed, by writing $\hat{S} = VX_w$, where V is an orthogonal matrix, one might observe that :

$$H(\hat{S}) = H(VX_w) = H(X_w) + \log_2 |V| = H(X_w) = \text{ cte.} \tag{2.17}$$

Hence, minimizing the mutual information for $\hat{S}$ is equivalent to minimize the sum of its marginal entropies, under the second order correlation constraint $\hat{S}\hat{S}^T = I_p$. Many work has been done to estimate a distribution entropy out of a sample. Some basic strategies try to estimate directly the distribution by using an histogram method. This is known as being highly dependent to the bin sized use to compute the histogram and therefore not used in practice. Kozachenko-Leonenko early proposed a binless entropy estimator, based on a k-nearest neighbor distances approach in 1987 [17]. It has been later improve [18][19]. Perhaps the most powerful mutual information estimator derived from this idea was proposed by Kraskov et al. in 2004 [20]. It is used by *sklearn*[2] that proposes a very efficient estimator, used in this internship. The effect of the number k used neighbors on the mutual information estimation is shown in figure 2.1.

Non-gaussianity

For a given mean and variance, the Gaussian distribution is the one with the highest entropy. This is the reason why we want to find the most non gaussian sources. A way to asses the gaussianity of a distribution, when having normalized moments of order 1 and 2, is to estimate its 4th order moment and compare it to the one expected by a gaussian. The kurtosis does this comparison. Given the probability density function, for a finite sample S, it is computed as :

$$K(S) = \frac{\mathbb{E}\left[\left(S - \overline{S}\right)^4\right]}{\mathbb{E}\left[\left(S - \overline{S}\right)^2\right]^2} - 3 \tag{2.18}$$

The constant 3 ensures that Gaussian signals have zero kurtosis while Super-Gaussian signals have positive kurtosis and Sub-Gaussian signals have negative kurtosis. Then ICA can be considered as an optimization process that maximizes the kurtosis amplitude, and make the extracted sources as non-normal as possible. Another proxy for non-gaussianity is negentropy. It measures the difference in entropy between a given distribution and the Gaussian distribution with the same mean and variance, by taking the Kullback-Leibler divergence between these two distributions. Thus, negentropy is always non-negative, is invariant by any linear invertible change of coordinates, and vanishes if and only if the signal is Gaussian. Using negentropy turns out to be more robust than kurtosis because less sensitive to outliers.

2.2.4 Various ICA heuristics

These numerous ways of interpreting and assessing the independence of a multivariate random variable led behind them even more ICA heuristics. Among some of them :

- The first ICA routine I implemented is named FOBI : Fourth Order Blind Identification. It was introduced by J. F. Cardoso in 1989 as the first step into ICA [21]. Based on data fourth order diagonalization, it is the natural extension of PCA. It is a very fast algebraic method. However it is not robust because very sensitive to outliers. This issue is clearly visible in figure 2.4 and led me to seek for other heuristics.
- An extension of FOBI was later proposed by J. F. Cardoso. It is based on a Joint Approximation Diagonalization of Eigen-matrices (JADE). It is still based on the fourth order analysis as it aims to maximize kurtosis. It is still in a non-gaussian approach of independence.
- FastICA is a very efficient and popular ICA method. Introduced by A. Hyvärinen in 2000 [22], it is based on a fixed-point iteration scheme that maximizes an approximation of non-Gaussianity. A very efficient Python routine is implemented on *sklearn* and was very useful for this internship.

2. More details here : https ://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_classif.html

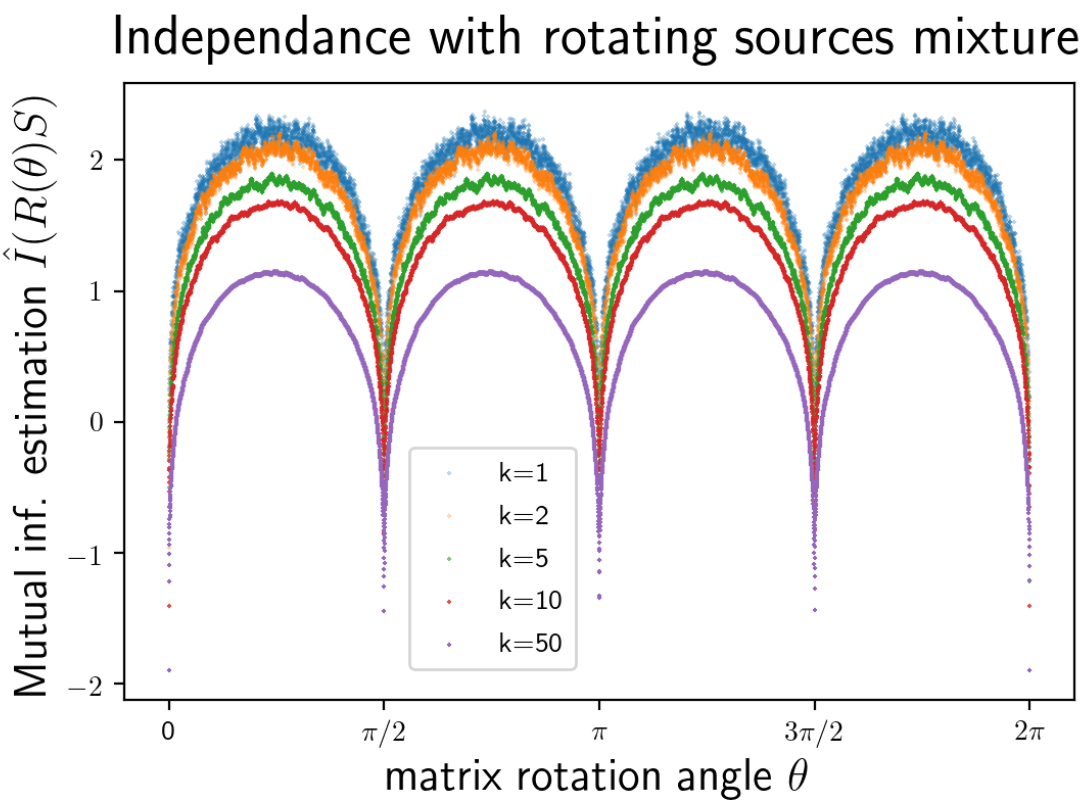


FIGURE 2.1 – Visualization of rotation influence $R(\theta)$ of 2D independent sources S over mutual information. The *sklearn* k-nearest neighbors mutual information estimator $\hat{I}_k$ is used. The estimator is strongly biased since mutual information should stay positive. This is not a problem in a minimization problem. The smoothness of the curve increases directly with k but also does the computation time. The data $R(\theta)S$ is always uncorrelated and independent if and only if $\theta \equiv 0[0, \pi/2]$. It was sampled from a very sharp distribution : mutual information is very sensitive when being close to independence.

2.2.5 My implementations

During this internship, I tried to implement my own ICA heuristics. My objective was to implement an independence optimization based method especially designed to our problem characteristics : we have few time observations $n \sim 50$ and matrix A should have a constant column, as shown in equation (2.5).

The first issue implies a perturbation in the covariance matrix estimated by time averaging, in practice we have only $SS^T \simeq I_l$. To solve this I implemented a method that takes both A and S as variables and relaxes the equality constraint $SS^T = I_l$ into a penalization term. It also allowed me to relax the constraint $X = AS$ into another penalization term. The problem formulation is :

$$\text{Solve} \underset{A \in \mathcal{M}_{p,l}(\mathbb{R}), S \in \mathcal{M}_{l,n}(\mathbb{R})}{\operatorname{argmin}} I(S) + \text{pen.}(SS^T - I_l) + \text{pen.}(X - AS) \tag{2.19}$$

Where I is a mutual information estimator. This was a total failure, the dimension of the search space is $(p+n) \times l \gg p \times l$ is way too high.

To solve the second issue, I implemented some penalization and lagragian routines for the constant column constraint in A. This was improving a bit results, but not enough. As trying to reformulate this constraint, a new idea came up : would it be possible to encode it directly in the subspace where optimization is performed ? This brought me to the concept of manifold optimization and its corresponding open-source Python package Pymanopt [3] [23]. The idea behind this is very simple : to solve an optimization problem that has a constraints that can be rephrased as belonging to a manifold, the research can be made

3. https ://www.pymanopt.org/

only on this manifold. The tricky part is to describe the manifold in such a way that it should allow classical optimization routines. This is the high added value of Pymanopt. However, I have not been able to finish the implementation a Python manifold class that would correspond to the various constraints I had. Nevertheless, the simplest strategy I started with turn out to be interesting enough to spend more time on it. After performing the PCA and getting the truncated X_w, ICA is reduced to assess the independence of VX_w where V could be any orthogonal matrix. Even better, because sources are defined up to a sign, we reduce the problem to the subspace of orthogonal matrices with a determinant of 1 : rotation matrices, $SO_l(\mathbb{R})$. It has a manifold structure, that is implemented in Pymanopt library. Hence, the new problem I proposed to solve is simply :

$$\text{Solve} \quad \underset{V \in SO_l(\mathbb{R})}{\operatorname{argmin}} \; I(VX_w) \tag{2.20}$$

Various optimization methods can be adopted with Pymanopt, but we explain the basic shared principle with figure 2.2.

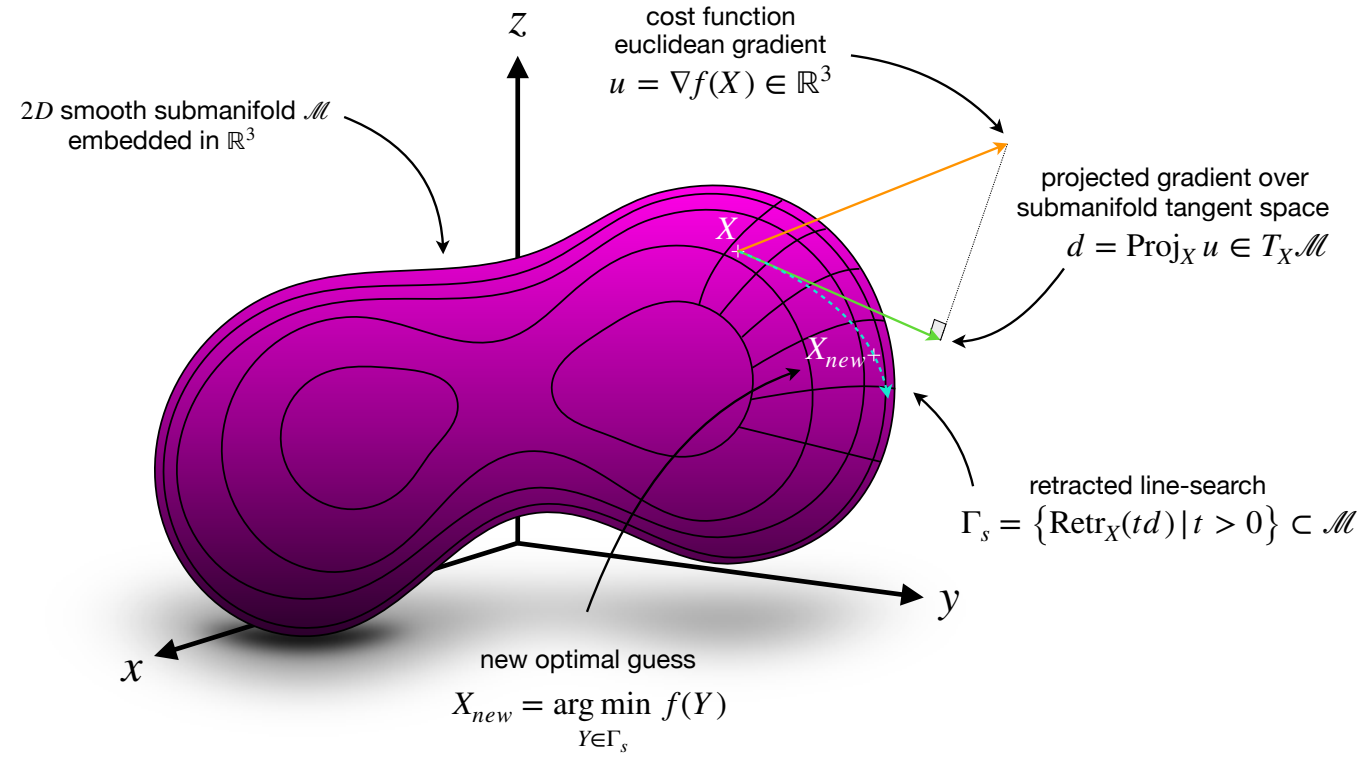


FIGURE 2.2 – Example of a 2D manifold $\mathcal{M}$ in purple embedded in a 3D space. Starting the optimization on a point X of $\mathcal{M}$, the gradient of the cost function is evaluated over the Euclidean space (orange) and then projected onto the tangent space $T_X\mathcal{M}$ (green) thanks to projection functions encoded in Pymanopt. Then a 1D optimization problem is defined over the retracted half-line (blue). Gradient properties ensure their exist a better solution X_{new} somewhere along this distorted hallf-line. Various solvers are proposed by Pymanopt such as Newton method.

2.2.6 Comparison using personal benchmarks

2D exemples have the advantage to be easily plotted and visualized. The 2D simulated data X was made first by choosing A, sources $S = (S_1, S_2)$ distribution and additive gaussian noise $\epsilon = (\epsilon_1, \epsilon_2)$ amplitude as follows :

$$X = AS + \epsilon \quad A = \begin{pmatrix} 1 & 0.1 \\ 1 & -0.2 \end{pmatrix} \quad S_{1,2} \underset{\textit{i.i.d.}}{\sim} \Gamma_{\pm}[k = 0.2, \theta = 1] \quad f_{\Gamma_{\pm}}(x, k, \theta) = \frac{|x|^{k-1} e^{-|x|/\theta}}{2\Gamma(k)\theta^k} \quad \epsilon_{1,2} \underset{\textit{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\epsilon)$$

A symmetric gamma density $\Gamma_{\pm}[k, \theta]$ was used for sources generation because of its easy control of the density shape. Choosing $k = 0.2$ sets a very sharp density, that fosters sources non-gaussianity and therefore the ICA convergence. A simulation of $n = 500$ samples is analyzed in figure 2.3. We performed 2 000 runs of these simulations and reported the corresponding error histograms in figure 2.4.

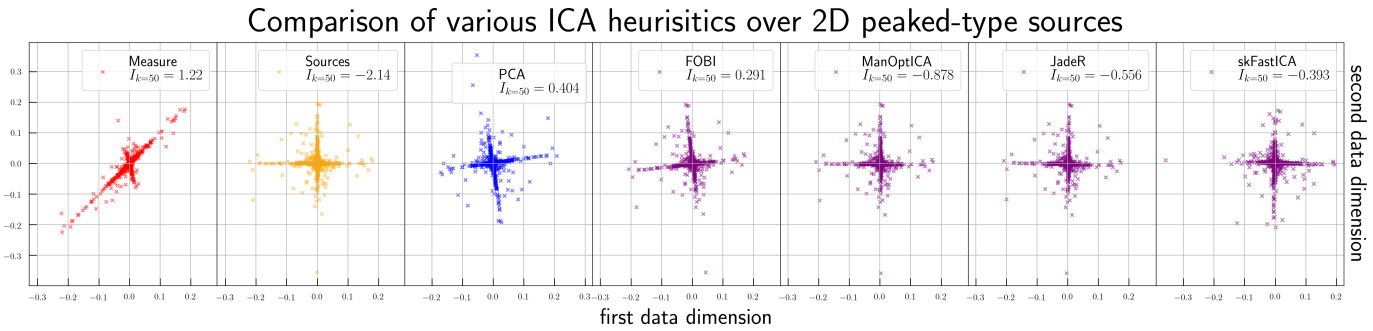


FIGURE 2.3 – 2D representation of the data. Independent sources in yellow were used to generate the measured data in red. The whiten data is in blue : orthogonal, but not aligned with horizontal and vertical axes (see figure 2.1 for the effect of rotation over independence). The 4 ICA heuristics are tested. FOBI is clearly not optimal. The 3 others fits perfectly, up to a permutation and ± 1 factor. My personal method ManOptICA stands with the best independence grade.

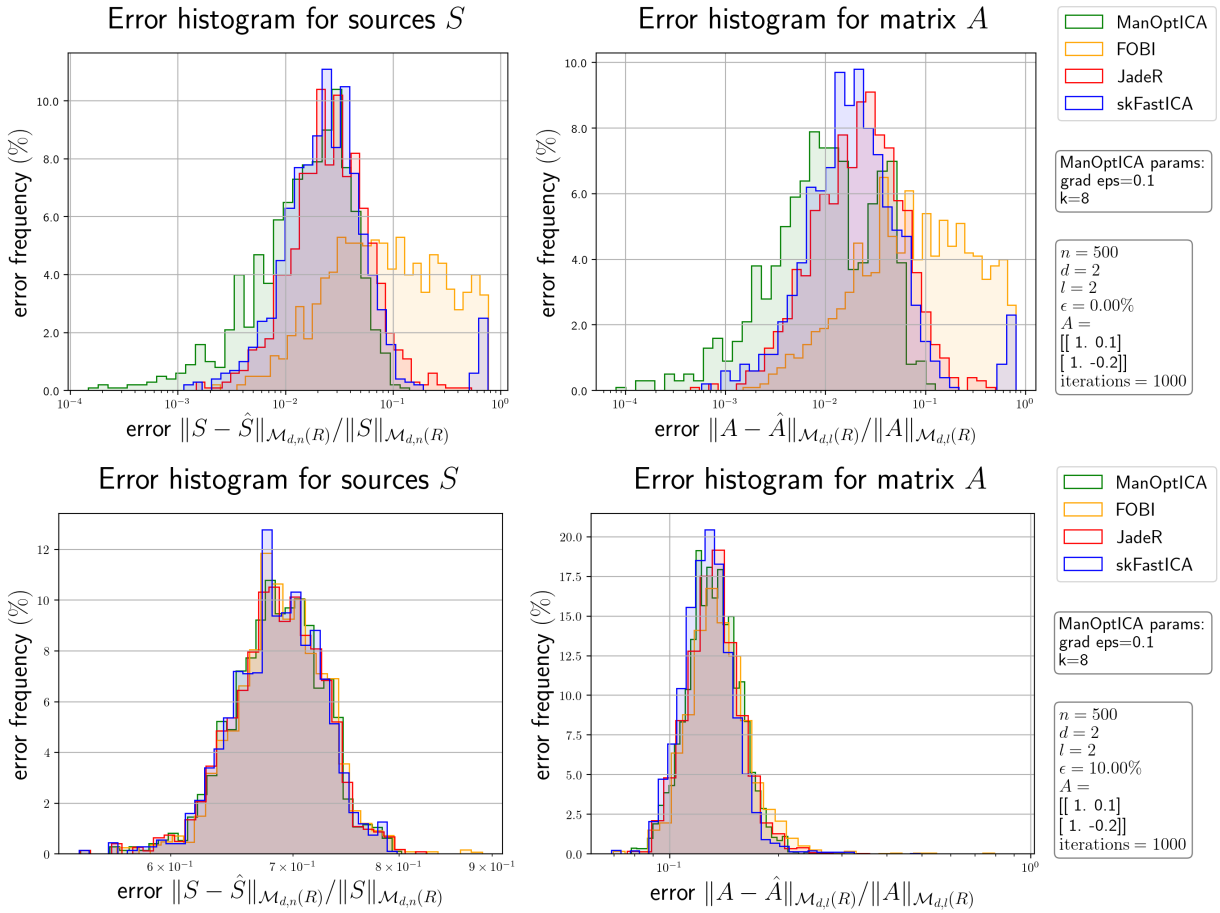


FIGURE 2.4 – Benchmark results in the 2D sharped case shown in figure 2.3. A was fixed while 2 000 samples of S were made. Half the input ICA data $X = AS$ was given without noise (top), the other half adding a 10% gaussian noise $X = AS + \epsilon$. ICA runs were made for each routine and sample X, their results $(\hat{A}, \hat{S})$ were compared to (A, S). The corresponding errors were stacked into histograms. Without noise, FOBI in yellow is not performing well, skFastICA and JadeR are equivalent up to some strong error bin for skFastICA, ManOptICA performs slightly better. With noise, all methods are equivalently not performing.

Chapitre 3

Application to exoplanet finding in HD189 and HD124

3.1 Introduction of the data

3.1.1 HD 189733 b

HD 189733 b is a gas giant blue color exoplanet that orbits a K-type star, 63.4 ± 0.9 light years away from us. Its mass is 1.162+0.058−0.039 Jupiters, it takes 2.2185733 ± 0.00002 days to complete one orbit of its star with a 0.0010 ± 0.0002 excentricity, and is 0.0313 AU from its star. It is a transiting planet with orbital an inclination of 85.76 ± 0.29°. Its discovery was announced in 2005 [24], by combining high-precision radial velocities that were measured with the ELODIE fiber-fed spectrograph on the 1.93 m telescope, and high-precision photometry that was obtained with the CCD Camera on the 1.20 m telescope, both at the Haute-Provence Observatory. It has been chosen to play the role of the control system : its stellar activity is weak. We extracted from the SOPHIE database the RV time series having 37 spectrum orders. We kept 55 instants, from 29/06/2007 to 07/09/2007.

3.1.2 HD12484 b

HD12484 b is a gas giant exoplanet that orbits a G-type star, 167 light-years from Earth. Its mass is 2.98 Jupiters, it takes 58.8 days to complete one orbit of its star with a 0.07 excentricity, and is 0.297 AU from its star. Discovered in 2016 by [25] with Radial Velocity method using SOPHIE. Its stellar activity is known to be important, this is the real test system. We extracted from the SOPHIE database 54 instants, from 17/09/2012 to 03/10/2018, for the RV time series having 37 spectrum orders.

3.1.3 Pre-processing

The first remark we can make is the poor number of time samples : we have only a fifty of instants. This raises the question of the precision and the robustness of time averaged estimations. Hopefully, the number of measured orders is high : 37. That will allow us to extract more 'meaningful' patterns during the PCA and lower the dimension to only significant signals. The results of PCA for the two planets are reported in figure 3.1.

As expected, the variances distribution of HD 189 is quite peaked onto its principal component, noted $PCA1$, that represents almost 90% of the full signal power while HD 124 is more diluted, having only 65% of its signal power gathered in its $PCA1$. These principal components $PCA1(t)$ are plotted for each planet in figures 3.2 and 3.3, along with the theoretical curves expected for the signal RV_{pl} induced by each planet and the averaged over orders AVG signals. We can see that $PCA1$ and AVG are very close from each other. This observation comforts us in the model used as it is a direct consequence of equation (2.5) that shows the specific structure of the expected mixing matrix A.

3.1.4 AVG and PCA1 : two proxies for RV extraction

As shown in figures 3.2 and 3.3, both AVG and $PCA1$ are close to the theoretical RV planet signal, with a respective SNR of 13.1 and 12.6 for HD 124 and 74 and 73 for HD 189. Hence, both can be used as proxy for the RV signal that has to be estimated. Indeed, the output of an ICA is $\hat{A}, \hat{S}$ with $\hat{S}$ containing l time series $(\hat{S}_1, \ldots, \hat{S}_l)$, making confusion to find the planet signal. Therefore, having such a free 'blind' proxy on hand is extremely valuable and allows us to find the planet induced signal

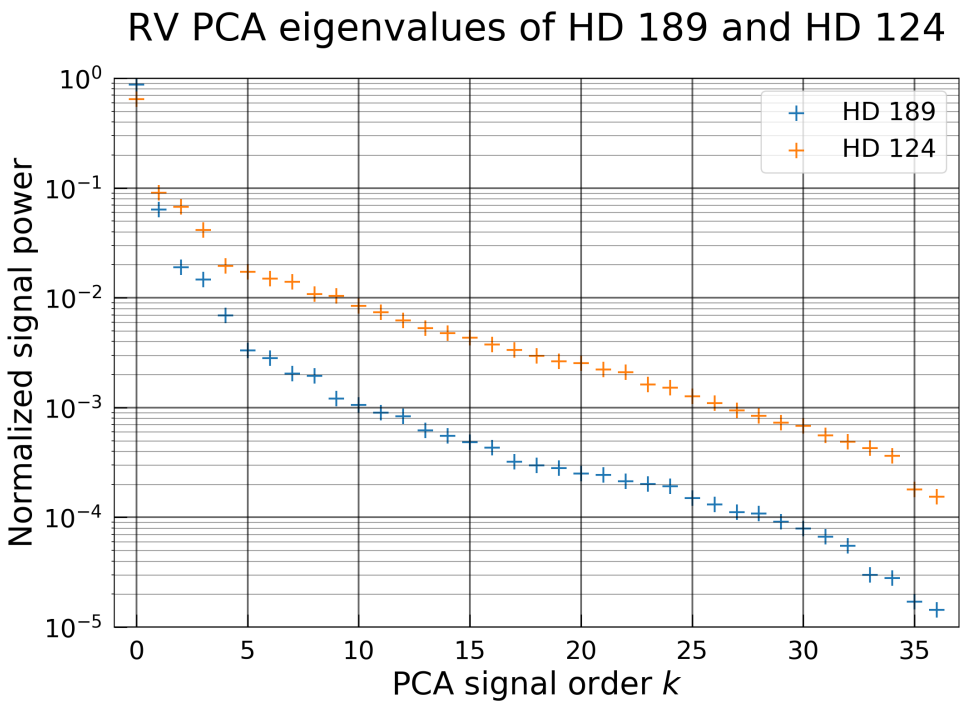


FIGURE 3.1 – Normalized distribution of eigenvalues stemming from the PCA of the RV data of HD 189 and HD 124. This can be seen as the temporal variance of each PCA component. HD 189 RV information is less diluted than HD 124 which is in accordance with the prior idea we had of these two systems that led us to choose them.

among all output sources. To find the best candidate for the planet signal among sources $(\hat{S}_1, \ldots, \hat{S}_l)$, we just select the more correlated one with our proxy :

$$\text{Find}(\hat{S}, RV_{proxy}) := \underset{RV_{signal} \in \{\hat{S}_1, \ldots, \hat{S}_l\}}{\text{argmax}} \left| \text{Cov}\left(RV_{signal}, RV_{proxy}\right) \right| \tag{3.1}$$

Of course, a scaling factor has to be used to recover entirely the ICA estimated planet signal out of the best estimated source. This is done in a 'blind' way, using the scaling factor correction :

$$\text{Find}(\hat{S}, RV_{proxy}) \frac{\text{Var}(RV_{proxy})}{\text{Cov}\left(\text{Find}(\hat{S}, RV_{proxy})\right)} \hookrightarrow \text{Find}(\hat{S}, RV_{proxy}) \tag{3.2}$$

3.1.5 Is their hope ?

If we throw the 37 PCA signals into the ICA, it will fail. Indeed the dimension is very high and freezes the convergence. But above all, now all the 37 signals will be put on an equal footing while only a few of them carry meaningful signature of the initial data. Hence it is necessary to lower the dimension according to the PCA results. But this has a cost : it increases the gap between the 'true' RV planet induced signal RV_{pl} and the closest estimation that can be found by ICA. We are allowed to use any linear combination of the input data to build an estimation of sources. Hence, a simple expression of the upper bound of the best signal candidate that could be found in the output of ICA, when given an input data $X_1, \ldots, X_r$ is :

$$\text{HopeLim}(X_1, \ldots, X_r) = \sup_{Y \in \text{Vect}(X_1, \ldots, X_r)} \text{SNR}(RV_{pl}, Y) \tag{3.3}$$

Different values of HopeLim are reported in table 3.1. It shows the decreasing best SNR performance we can get using ICA when reducing the input data dimension. However, this reduction of hope depends on the reduction choice. Two methods have been compared : basic PCA truncation, as motivated in the second part, and orders grouping. The latter method, that I implemented, consists in clustering the RV orders into r groups where r is the targeted reduced dimension, and return, by group, the averaged RV over orders. This is a very basic reduction that also appears naturally during the measuring process. Indeed, the CCD screen does not catch one light order per raw of pixels as seen in figure 1.2. The physical order is naturally spread over several raws.

The measured signal over these raws has to be grouped and averaged before building a final digitized version of this order. This is done by the SOPHIE processing pipeline that converts a 2154×4102 pixels signals into a FITS file containing 40 orders (each order is a 4077 1D array). Of course, in practice we do not have access to RV_{pl} and thus, this upper bound for the best SNR that could be expected is not known. The performance of the ICA can drastically drop if the dimension reduction is not optimal.

reduction dim.	37	6	4	2
HD 124 – PCA trunc.	146	16	16	14
HD 124 – Order Group.	146	27	29	21
HD 189 – PCA trunc.	525	140	114	93
HD 189 – Order Group.	525	140	135	124

TABLE 3.1 – HopeLim results with dimension reduction

3.2 Stochastic ICA pipeline

3.2.1 The stochastic approach using post-processing

To thwart this issue, I implemented a stochastic ICA pipeline, that runs various ICA trials with various parameters, grades blindly the result using a RV proxy such that AVG or PCA1, and aggregates the results according to these grades. Because this will need many ICA runs, I chose to use the FastICA routine implemented in *sklearn*. The parameters I proposed to use in the stochastic approach are the following :

- Reduction : the reduction method called. We saw it could be for instance orders grouping or PCA truncating.
- l : the number of sources to identify. This is the dimension that will be targeted when calling the Reduction method.
- A_{init} : the initial orthogonal matrix used to start independence optimization. It operates directly on the whitened data with dimension l.
- ϵ_{add} : the added noise to initial given X data. Useful to help PCA convergence if X^T is not full rank.

An ICA run could be resumed as follow :

$$\hat{S}(\text{Reduction}, l, A_{init}, \epsilon_{add}) := \text{FastICA}\left(\text{Reduction}(X + \epsilon_{add}, l), A_{init}\right) \tag{3.4}$$

Throwing N ICA runs for various values of these parameters will raise N sources estimations $\left\{\hat{S}^{(1)}, \ldots, \hat{S}^{(N)}\right\}$. As we changed l over the trials, two different runs $\hat{S}^{(i)}, \hat{S}^{(j)}$ may have different dimensions $l_i \neq l_j$. This is not an issue since we have a proxy for RV and a Find method to find the best candidate in each ICA run. The only missing part is how to assess a run. If we assume to handle this step, the post-processing pipeline can be resumed in this equation :

$$\widehat{RV}_{pl} = \text{PostProcess}\left(\left\{\hat{S}^{(1)}, \ldots, \hat{S}^{(N)}\right\}, RV_{proxy}\right) := \frac{1}{h}\sum_{j=1}^{N} \text{Grade}\left(\text{Find}(\hat{S}^{(j)}, RV_{proxy}), RV_{proxy}\right) \text{Find}(\hat{S}^{(j)}, RV_{proxy}) \tag{3.5}$$

where :

$$h := \sum_{j=1}^{N} \text{Grade}\left(\text{Find}(\hat{S}^{(j)}, RV_{proxy}), RV_{proxy}\right)$$

3.2.2 Grading an ICA output

How to give a grade to an ICA run $\hat{S}$? They are many ways to answer this question, more or less sophisticated. The one I found to be the most efficient is actually very simple and straightforward. We have a proxy for the signal we want to find in sources $\hat{S}$, let us simply make sure it is enough correlated (up to the sign) to the signal we found :

$$\text{Grade}(RV_{found}, RV_{proxy}) = \begin{cases} 1 & \text{if } \left|\text{ Cov }\left(RV_{signal}, RV_{proxy}\right)\right| > \eta \\ 0 & \text{otherwise.} \end{cases} \tag{3.6}$$

This is basically a ceiling function that drops out the fails of ICA. The acceptance level η was set to 0.95. One might wonder : to what extent is this grading depending on the proxy used ? Is it even fair to grade our result using a proxy of the solution ? First, the proxy used and proposed in this project are blind : averaging over orders or taking the principal component are general routines that are not asking any information about the system observed. Later, one might wonder if this post-processing scheme

is biased by the proxy used in the grading system. Why would the estimation converge towards RV_{pl} rather than RV_{proxy} ? This is an important question. My understanding is that ICA routine is unbiased and should converge towards the expected solution in general. But it is also free to deliver any kind of output if not well pre-processed. Hence, applying this post-selection is, in a sense, equivalent to bounding the search space. In practice, this grading system turns out to be efficient for 3 reasons :

— using either RV_{pl} or AVG or PCA1 as a proxy for RV_{pl} does not change the SNR of the post-processing result $\widehat{RV}_{pl}$. In that sense, this grading system is few dependent on the proxy its uses.
— it is indeed discriminating very well ICA failures. Thus, only interesting values of the parameters are kept with a good grade. This enables us to access to these interesting parameters like the number of sources l to find in the signal.
— it straightforward to compute. This indicator is cheap and can be neglected compared to the computational time of an ICA run.

3.3 Results and discussion

3.3.1 HD 124

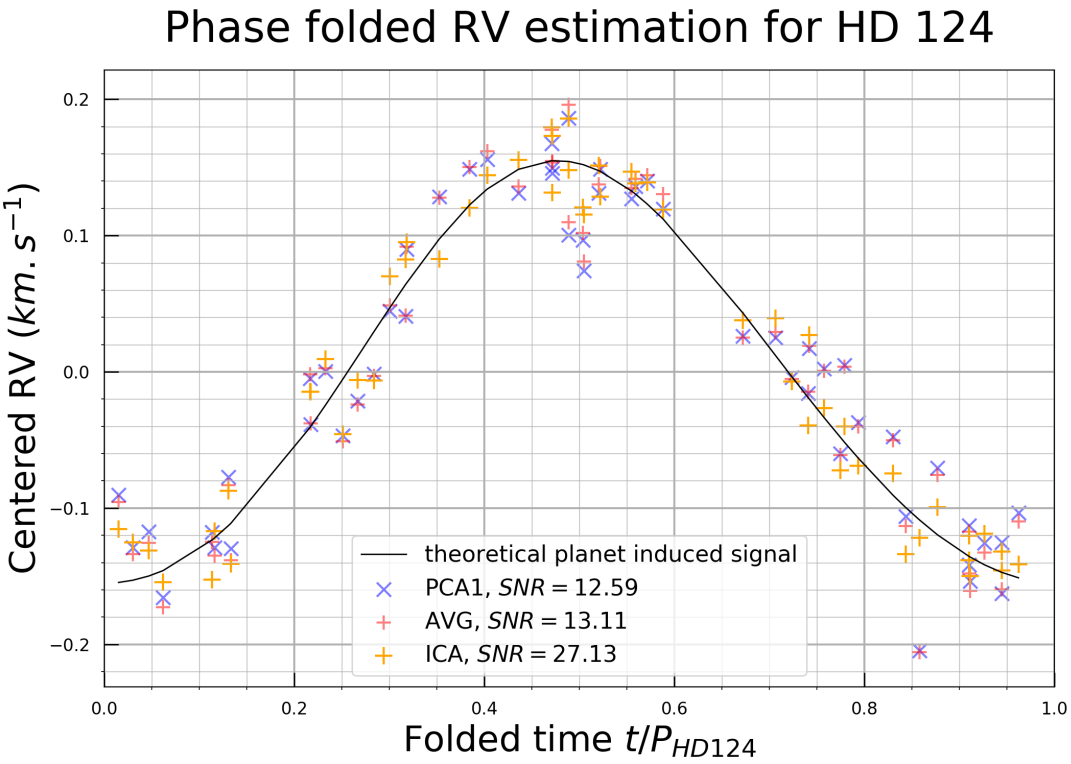


FIGURE 3.2 – Final results applied to HD 124 : stochastic ICA combined with post processing using AVG as the RV proxy, and an acceptance level $\eta = 0.95$. Data dimension is reduced with averaging over groups, all dimensions $2 \leq l \leq 7$ where equally targeted.

The PCA of HD 124 in figure 3.1 shows us 4 main components. When throwing our stochastic ICA heuristic, with a reduction based on averaging orders into $2 \leq l \leq 7$ dimensions, we obtain the final guess $\widehat{RV}_{pl}$ with a SNR of 27.3, plotted in figure 3.2. In comparison, AVG and PCA1 had a SNR of 13. The SNRs of stochastic ICA thrown, each one over a specific dimension $l \in [\![2, 7]\!]$, are reported in table 3.2. We see that the best estimation is obtained for $l = 4$. This is in accordance with table 3.1 : reducing the dimension to 4 signals using a simple averaging over 4 groups of orders decreases the SNR HopeLim only to 29. This seems to be a perfect trade-off between reducing consequently the dimension and keeping enough diversity to build a close estimation, that can only stay in the linear space spanned by these 4 signals. Of course, in the general case, we are not supposed

to have the information of table 3.1 and 3.2. But we can say here that our post-processed stochastic ICA heuristic turns out to be powerful, since it has more than doubled the SNR, without any prior information about the signal.

sourc. dim. l	[2 :7]	2	3	4	5	6	7
SNR	27.3	20.3	18.3	25.0	21.5	19.1	21.2

TABLE 3.2 – HD 124 post-processing results

3.3.2 HD 189

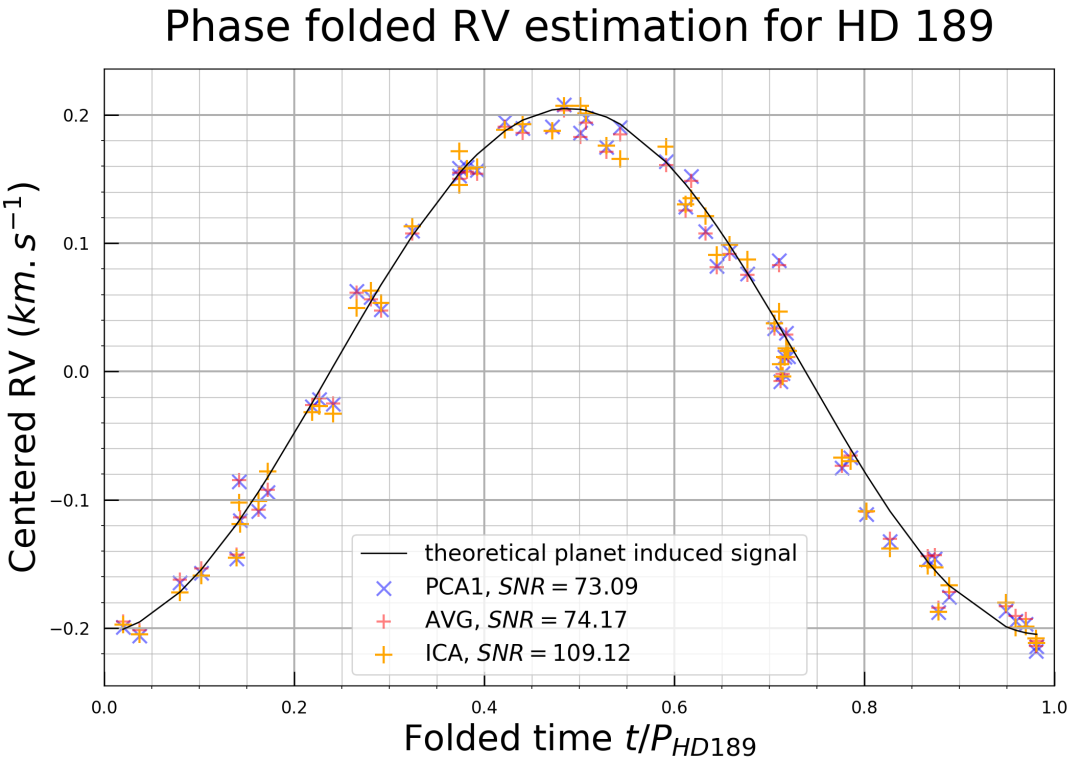


FIGURE 3.3 – Final results applied to HD 189 : stochastic ICA combined with post processing using AVG as the RV proxy, and an acceptance level $\eta = 0.95$. Data dimension is reduced with averaging over groups, all dimensions $2 \leq l \leq 5$ where equally targeted.

sourc. dim. l	[2 :5]	2	3	4	5
SNR	109	110	36	73	44

TABLE 3.3 – HD 189 post-processing results

The PCA of HD 189 in figure 3.1 reveals 2 strong components followed by 2 medium ones. When throwing our stochastic ICA heuristic, with a reduction based on averaging orders into $2 \leq l \leq 5$ dimensions, we obtain the final guess $\widehat{RV}_{pl}$ with a SNR of 109, plotted in figure 3.3. In comparison, AVG and PCA1 had a SNR of 74. The SNRs of stochastic ICA thrown, each one over a specific dimension $l \in [\![2, 5]\!]$, are reported in table 3.3. We see that the best estimation is obtained for $l = 2$. This is quite in accordance with table 3.1 : reducing the dimension to 2 signals using a simple averaging over 2 groups of orders decreases the SNR HopeLim only to 124. The trade-off between reducing consequently the dimension and keeping enough diversity seems to

be harder to find than in the previous case. Table 3.3 shows that it would have been better to throw the ICA only targeting $l = 2$: the SNR would be 110. Of course we are not supposed to have this information, but we did not use it and did not need it : the overall SNR is 109. Our post-processing heuristic demonstrates a great robustness : even tough trying ICAs with meaningless signals, the corrupted outputs have been dropped by the post-processing selection that caught only good estimations. However, in this case, ICA does not perform as well as the previous one, in terms of SNR augmentation.

Conclusion and personal assessment

This internship was the occasion for me to focus my attention on a specific project for few months. Exploring the bibliography of a new field, implementing some Python routines and looking for mathematical explanations is very similar to many school projects I have already experienced. Obviously I was expecting more from this internship, such as working in a lab and multiple my exchanges with researchers, may be discover another approach of science, a new culture and a new country. Nevertheless, this experience was one of the first opportunity where I really took the time to immerse deeply myself into a problem and to develop my own ideas. As a habit, I took the mission proposed by Rodrigo and Alain as a personal challenge. But for once, time was on my side.

I have been very interested by the problem I faced. Analyzing the radial-velocities of HD 189 and HD 124 was exciting. This few $2 \times 40 \times 60$ data points are the precious result of tens light years travel of a signal, captured by a large telescope and analyzed by an extremely sensitive spectrograph, carefully collected and reported by astronomers over years. The precision reached by the spectrograph is very high, but does not supersede the need for theoretical analysis of the observed phenomena. I deeply enjoyed working on a subject that involved simultaneously an experimental and a theoretical challenge. This duality is, according to me, a strength of astronomy.

The mathematical dimension of the project was particularly adapted to my knowledge level. Sufficiently accessible and complex enough, it was a great trade-off that allowed me to bring new ideas and seek for their potential. However, it was hard for me to take a step back and have the distance needed to judge the prospects of success of the various paths I entered. Of course, many of them were dead ends. More than once, unwillingly, I rushed the bibliography step and found myself wasting time reinventing the wheel. It has its pedagogical advantage but still, if I had to do this project again, I would intentionally spend much more time on the bibliographic research, until I am almost certain not to miss any important concept, both in astrophysics and applied mathematics.

Finally, the principal lesson I will learn from this experience is the power of the questioning process. I have been hit by the importance of questions in my progress. Many times, asking the right question would unlock the problem. Not by solving it, but rephrasing it, deleting the superfluous complications and enhancing the central problem, exposed in its simplest form. It seems that half the solution resides in the problem formulation. This idea could not be more actual when applied to the optimization problem hidden in ICA. The formulation includes directly the objective function, the variables at stake, their constraints and the set were they should evolve. Simply encoding a constraint into this set allowed me to suggest a new ICA routine. The only limit to this principle is time. As the research process goes undoubtedly forward in time, the number of questions raised grows faster than the number of answers found. And so is the mean feeling of incompleteness.

"L'astronomie est la seule science où
nous ne pouvons pas faire d'expériences :
nous ne pouvons ni recréer le big bang en laboratoire
ni concocter des étoiles dans des éprouvettes
Alors comment connaître l'univers ?
La lumière vient à notre secours.
Elle est le messager du cosmos par excellence.
Elle est ma compagne.
C'est elle qui me permet de communiquer avec le cosmos et de l'étudier.
C'est elle qui véhicule les fragments de musique et les notes éparses
de la mélodie secrète de l'univers que l'homme tente de reconstituer."

Trinh Xuan Thuan, *Le Cosmos et le Lotus*, 2011

Table des figures

Bibliographie

[1] GRAVITY Collaboration, S. Lacour, M. Nowak, J. Wang, O. Pfuhl, F. Eisenhauer, R. Abuter, A. Amorim, N. Anugu, M. Benisty, J. P. Berger, H. Beust, N. Blind, M. Bonnefoy, H. Bonnet, P. Bourget, W. Brandner, A. Buron, C. Collin, B. Charnay, F. Chapron, Y. Clénet, V. Coudé du Foresto, P. T. de Zeeuw, C. Deen, R. Dembet, J. Dexter, G. Duvert, A. Eckart, N. M. Förster Schreiber, P. Fédou, P. Garcia, R. Garcia Lopez, F. Gao, E. Gendron, R. Genzel, S. Gillessen, P. Gordo, A. Greenbaum, M. Habibi, X. Haubois, F. Haußmann, Th. Henning, S. Hippler, M. Horrobin, Z. Hubert, A. Jimenez Rosales, L. Jocou, S. Kendrew, P. Kervella, J. Kolb, A.-M. Lagrange, V. Lapeyrère, J.-B. Le Bouquin, P. Léna, M. Lippa, R. Lenzen, A.-L. Maire, P. Mollière, T. Ott, T. Paumard, K. Perraut, G. Perrin, L. Pueyo, S. Rabien, A. Ramírez, C. Rau, G. Rodríguez-Coira, G. Rousset, J. Sanchez-Bermudez, S. Scheithauer, N. Schuhler, O. Straub, C. Straubmeier, E. Sturm, L. J. Tacconi, F. Vincent, E. F. van Dishoeck, S. von Fellenberg, I. Wank, I. Waisberg, F. Widmann, E. Wieprecht, M. Wiest, E. Wiezorrek, J. Woillez, S. Yazici, D. Ziegler, and G. Zins. First direct detection of an exoplanet by optical interferometry : Astrometry and K -band spectroscopy of HR 8799 e. *Astronomy & Astrophysics*, 623 :L11, March 2019.

[2] Rodrigo F. Díaz. Modelling Light and Velocity Curves of Exoplanet Hosts. In Tiago L. Campante, Nuno C. Santos, and Mário J. P. F. G. Monteiro, editors, *Asteroseismology and Exoplanets : Listening to the Stars and Searching for New Worlds*, volume 49, pages 199–224. Springer International Publishing, Cham, 2018. Series Title : Astrophysics and Space Science Proceedings.

[3] Adam S. Burrows. Highlights in the study of exoplanet atmospheres. *Nature*, 513(7518) :345—352, 2014.

[4] Michel Mayor, Christophe Lovis, and Nuno C. Santos. Doppler spectroscopy as a path to the detection of Earth-like planets. *Nature*, 513(7518) :328–335, September 2014.

[5] Michel Mayor and Didier Queloz. A jupiter-mass companion to a solar-type star. *Nature*, 378(6555) :355—359, 1995.

[6] M. Zechmeister, A. Reiners, P. J. Amado, M. Azzaro, F. F. Bauer, V. J. S. Béjar, J. A. Caballero, E. W. Guenther, H.-J. Hagen, S. V. Jeffers, A. Kaminski, M. Kürster, R. Launhardt, D. Montes, J. C. Morales, A. Quirrenbach, S. Reffert, I. Ribas, W. Seifert, L. Tal-Or, and V. Wolthoff. Spectrum radial velocity analyser (SERVAL). High-precision radial velocities and two alternative spectral indicators. *Astronomy & Astrophysics*, 609 :A12, January 2018. arXiv : 1710.10114.

[7] Ian Czekala, Kaisey S. Mandel, Sean M. Andrews, Jason A. Dittmann, Sujit K. Ghosh, Benjamin T. Montet, and Elisabeth R. Newton. Disentangling Time-series Spectra with Gaussian Processes : Applications to Radial Velocity Analysis. *The Astrophysical Journal*, 840(1) :49, May 2017.

[8] F. Bouchy, F. Pepe, and D. Queloz. Fundamental photon noise limit to radial velocity measurements. *Astronomy & Astrophysics*, 374(2) :733–739, August 2001.

[9] Trifon Trifonov, Lev Tal-Or, Mathias Zechmeister, Adrian Kaminski, Shay Zucker, and Tsevi Mazeh. Public HARPS radial velocity database corrected for systematic errors. *Astronomy & Astrophysics*, 636 :A74, April 2020.

[10] Xavier Dumusque. Radial Velocity Surveys, July 2015.

[11] C. Lovis and D. Fischer. *Radial Velocity Techniques for Exoplanets*, pages 27–53. Seager, S., 2010.

[12] Brett M. Morris, H. Jens Hoeijmakers, Daniel Kitzmann, and Brice-Olivier Demory. Hunt for Starspots in HARPS Spectra of G and K Stars. *The Astronomical Journal*, 160(1) :5, June 2020.

[13] H. Korhonen, J. M. Andersen, N. Piskunov, T. Hackman, D. Juncher, S. P. Järvinen, and U. G. Jørgensen. Stellar activity as noise in exoplanet detection – I. Methods and application to solar-like stars and activity cycles. *Monthly Notices of the Royal Astronomical Society*, 448(4) :3038–3052, April 2015.

[14] X. Dumusque, I. Boisse, and N. C. Santos. SOAP 2.0 : A tool to estimate the photometric and radial velocity variations induced by stellar spots and plages. *The Astrophysical Journal*, 796(2) :132, November 2014. arXiv : 1409.3594.

[15] Jonathon Shlens. A Tutorial on Independent Component Analysis. *arXiv :1404.2986 [cs, stat]*, April 2014. arXiv : 1404.2986.

[16] Jean-François Cardoso. Dependence, Correlation and Gaussianity in Independent Component Analysis. *Journal of Machine Learning Research*, 4(7-8) :1177–1203, 2004. Place : US Publisher : MIT Press.

[17] Leonenko N. N. Kozachenko F. J. Sample Estimate of the Entropy of a˜Random Vector. *Probl. Peredachi Inf.*, pages 9–16, 1987.

[18] Jonathan D. Victor. Binless strategies for estimation of information from neural data. *Physical Review E*, 66(5) :051903, November 2002.

[19] Damiano Lombardi and Sanjay Pant. A non-parametric k-nearest neighbor entropy estimator. *Physical Review E* , January 2016.

[20] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6) :066138, June 2004.

[21] Jean-François Cardoso. Source separation using higher order moments. In *in Proc. ICASSP*, pages 2109–2112, 1989.

[22] A. Hyvärinen and E. Oja. Independent component analysis : algorithms and applications. *Neural Networks*, 13(4-5) :411–430, June 2000.

[23] James Townsend, Niklas Koep, and Sebastian Weichwald. Pymanopt : A python toolbox for optimization on manifolds using automatic differentiation. *Journal of Machine Learning Research*, 17(137) :1–5, 2016.

[24] F. Bouchy, S. Udry, M. Mayor, C. Moutou, F. Pont, N. Iribarne, R. Da Silva, S. Ilovaisky, D. Queloz, N. C. Santos, D. Ségransan, and S. Zucker. ELODIE metallicity-biased search for transiting Hot Jupiters : II. A very hot Jupiter transiting the bright K star HD 189733. *Astronomy & Astrophysics*, 444(1) :L15–L19, December 2005.

[25] G. Hébrard, L. Arnold, T. Forveille, A. C. M. Correia, J. Laskar, X. Bonfils, I. Boisse, R. F. Díaz, J. Hagelberg, J. Sahlmann, N. C. Santos, N. Astudillo-Defru, S. Borgniet, F. Bouchy, V. Bourrier, B. Courcol, X. Delfosse, M. Deleuil, O. Demangeon, D. Ehrenreich, J. Gregorio, N. Jovanovic, O. Labrevoir, A.-M. Lagrange, C. Lovis, J. Lozi, C. Moutou, G. Montagnier, F. Pepe, J. Rey, A. Santerne, D. Ségransan, S. Udry, M. Vanhuysse, A. Vigan, and P. A. Wilson. The SOPHIE search for northern extrasolar planets : X. Detection and characterization of giant planets by the dozen⋆⋆⋆. *Astronomy & Astrophysics*, 588 :A145, April 2016.