UNIVERSITÉ DE GENÈVE

FACULTÉ DES SCIENCES

Master report : Mitigating stellar activity in radial velocities

Improvement towards the lightest planets searching quest

Michael Cretignier

University of Geneva

June 6, 2018

ABSTRACT

We investigated the radial velocities (RV) of each spectral line in HARPS and HARPS-N spectra to see if some of them are more sensitive or not to stellar activity. We used as test case the K dwarf star α Cen B exhibiting, in the 2010 radial velocity data from HARPS, a strong modulation due to stellar activity, with peak-to-peak variations of 10 m/s. By carefully selecting the spectral lines, we can mitigate stellar activity effects and bring the rms down to a level of 77 cm/s. Machine learning algorithms have been used to increase the number of lines, that are not correlated with activity, based on their physical properties. The same protocol has then been applied for 2009 and 2011 α Cen B's data and for our own star : the Sun. If our method successfully mitigates stellar activity on stellar rotational timescale for both stars, the particular long and well sampled data of the Sun, distributed over 2 years, show that also long-period magnetic cycle contamination can be reduced from -3.57 m/s/year down to 0.18 m/s/year. The activity signal in radial velocity according to our reduction is strongly correlated with the $\log R'_{HK}$ index without being utterly explained by it. It confirms that our method is mitigating the activity and that the $\log R'_{HK}$ cannot be used to correct efficiently the stellar activity signal. As a final point, simulations of planetary detection have been performed, injecting planetary signals in the 2-years RV data of the Sun. With our best selection of lines unaffected by activity, the detection limits can be improved down to a level of 40 cm/s, compared to 60 cm/s when using all the spectral lines. Also, planets with orbital periods close to the rotational period are now less affected by activity and allow easier planetary detections. We note however, that our best selection of unaffected lines only contain 10% of the RV information present in HARPS or HARPS-N spectra. Thus the photon noise on the RVs derived by such a line selection is increased by a factor of 3. This is not a problem for stars where the photon noise is not limiting, however we still have to investigate the effect on stars with a low SNR. Although it seems that this factor of 3 will kill such type of analysis on star with low SNR, it is not straightforward as reducing strongly the correlated noise induced by activity can strongly counter balance the increase of white photon noise.

Key words. radial velocity contamination – HARPS – alpha Centaurus – Sun – machine learning algorithm – atomic physics of the line

Contents

1. Introduction

With the advent of the next generation of spectrograph, the question of stars' activity is becoming more and more a matter of concern. Indeed, stellar activity signals are known to provide false positive in planetary radial velocity (RV) detection. And in the quest of habitable worlds, spots and active regions can produce a signal with amplitude well beyond the RV signal of an Earth-like planet. The oldest measurement of star activity are likely observations of sunspots, as sketched for instance in XVIIth century (see Fig.1), but older observations are already drawn by Galilei and even before him by John of Worcester in 1182. Because of their abundance and since the Sun is the only star optically resolved, after decades of studies and data accumulation, some rules of magnetic cycle (Hathaway 2015) have been understood, like the 11 years solar Schwabe cycle but also longer and shorter variations, of a few minutes until millenium (Usoskin et al. 2007). For other stars, the resolutions of the telescopes are still not sufficient to optically resolve the active regions and only their impact on the lines are observable without being completely understood.

In section 2, we briefly explain what are the stellar physical processes perturbing the RV measurements. The classical method of RV determination as well as the new method used for this work are then presented in section 3. The main idea being to find out a way to identify and separate the lines affected by the stellar activity from those which are not. We investigated two methods applied on the special case of α Centaurus B. The first is based on flux variation and will be presented in section 4. The second method, detailed in section 5, is based on the analysis of the RV of each spectral line. Then depending if those lines are affected or not by activity, we can group them to increase the RV precision, and therefore get useful RVs affected or not by stellar activity. We also investigated in this section the use of Machine learning algorithms to select affected and not affected lines based on their physical properties. Because we found the second method sucessful, we applied it to our own star, the Sun, as described in section 6. In section 7, we provide evidences that the contamination removed is genuinely due to the activity of the stars. Further investigations are made, in section 8, between the lines affected and unaffected for the two stars studied, particularly to see if the selections of both stars are exchangeable or if they present some overlap. Finally, simulations of planetary detections are performed to evaluate the improvement made by our work in section 9 and we conclude in section 10.

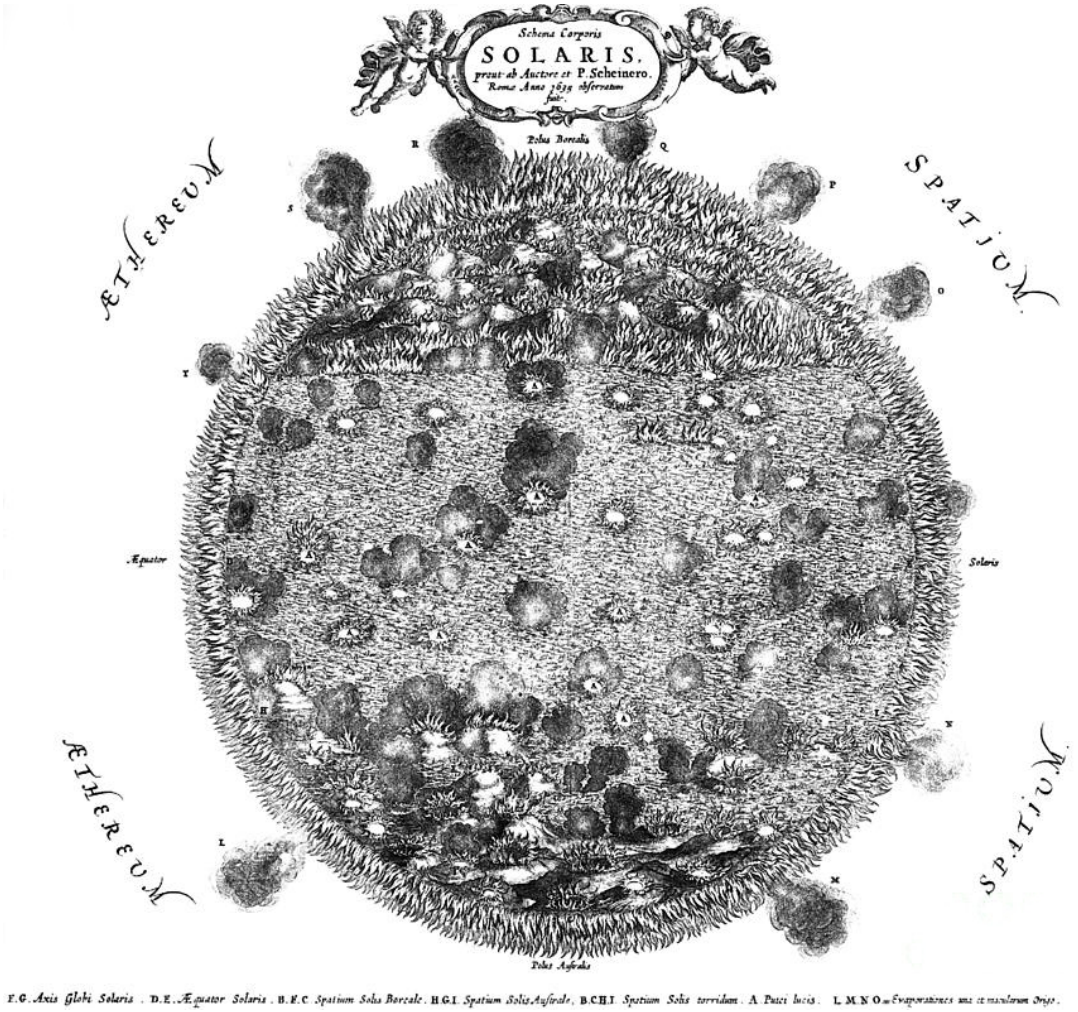

Fig. 1: Sketch of the Sun and its spots made in 1664 by Athanasius Kircher. Sunspots were intepret as the ashes of active volcanoes. (*Crédits : https://sciencesource.com/archive/Schema-Corporis-Solaris–Mundus-Subterraneus–17th-Century-SS2494633.html*)

2. Stellar impact on the radial velocities

For sufficiently cold temperature, more complex molecules and ions are formed in the stellar atmospheres. As a direct consequence, in the external layers of these stars, the opacity increases and the star needs to evacuate its energy providing other mechanisms of energy transport such as convection to maintain the energy flux. Because the opacity is closely related to the ionisation state of the matter, the core of solar type star is radiative, whereas the envelope is convective mainly because of the opacity caused by the H^- ion. For cool stars like late M-dwarfs, even the radiative core can disappear and the star is fully convective.

In the mixing length theory, convection is simplified has the movement of «hot bubbles» riding toward the surface and cold bubbles falling toward the center. The upflows appear, thus, as blueshifted and the downflows redshifted. The effect of convection is to produce a net blueshifted contribution. Indeed, since the hot upflows are by definition hotter than the cool ones, they emit more blueshifted photons than the cold downflows emitting redshifted photons. Secondly, the convection patterns of hot matter are more extended than the cold downflows. As an example, we display in Fig.2 and Fig.3 the magneto-hydrodynamics simulations of convective velocities from Beeck et al. (2013a) for 4 different types stars. The Fig.2 is the convective pattern of the vertical velocity from layers in average around the optical surface. Hot blueshifted convective cells are more extended than the infalling cold matter. This is even more true for deeper layer (Fig.3) where convection cells are even larger. Let us remark that, in order to preserve mass continuity, the cold matter is falling faster. In the same paper, the authors fixed at 5.1, 2.0, 1.1 and 0.72 km/s the vertical rms velocities at the optical surface for the 4 stars (respectively F3V, G2V, K0V and M0V stars). Such decrease is coherent since hotter stars have more energy to evacuate and less available convective material, therefore the speed must increase to compensate.

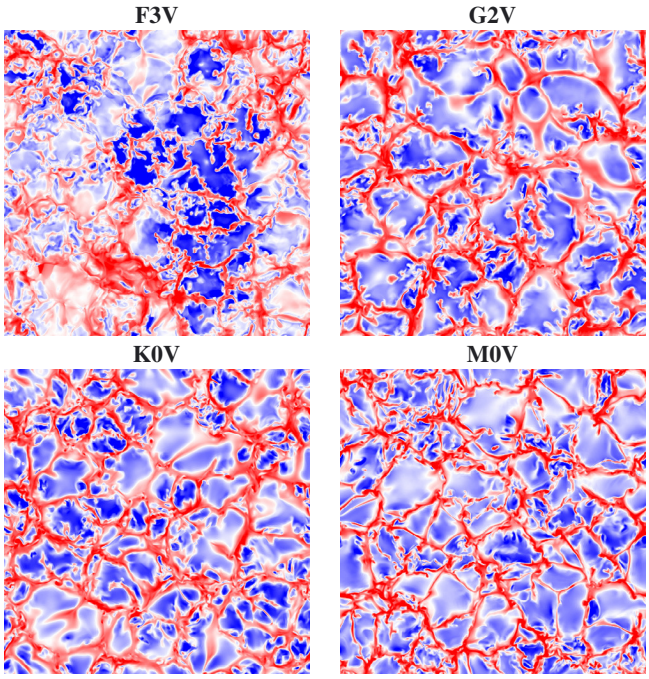

Fig. 2: Simulations of the vertical velocity of convection flows for 4 different type stars at a geometrical depth level around the optical surface ($\tau_R = 1$). Hot matter, blueshifted since moving toward the surface (blue), is forming more extended pattern ($\sim$ 55% of the area) than the infalling cold matter (red). As a comment, the color code and scales are not the same for the 4 stars. (*Credits : Beeck et al. (2013a)*)

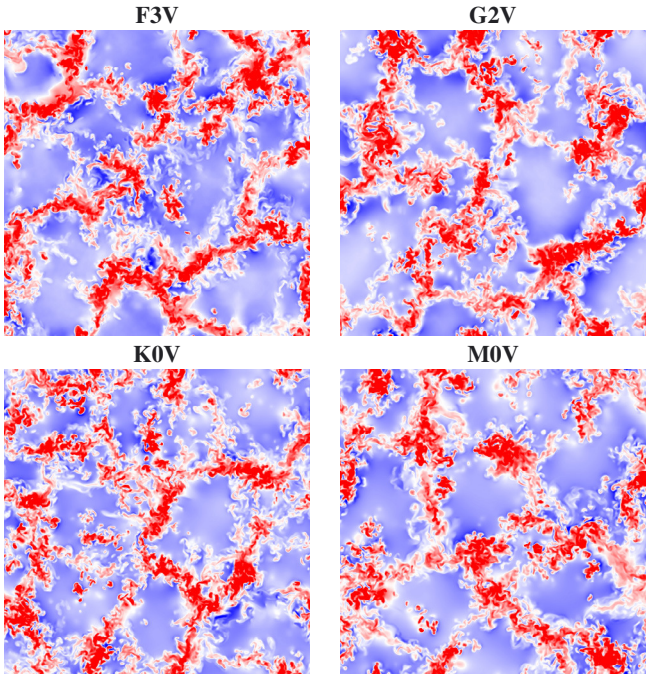

Fig. 3: Same as Fig.2 for deeper layers (4.6 times heights pressure scale deeper than the optical surface). The blueshifted convection pattern is larger ($\sim$ 65% of the area) in comparison with the optical surface level Fig.2. To maintain the mass conservation, the infalling matter is moving faster than the upflows : the color is strongly red against pale blue for the hot matter. (*Credits : Beeck et al. (2013a)*)

For the Sun, an order of magnitude for the velocity is 1 km/s with a peak of 2.6 km/s in the profile of the velocity convection at $\sim 10000°K$ (Maeder 2009). Near the surface, the average velocity of granulation is of the order of hundred of m/s (Meunier et al. 2017). In addition to these movements of convection, strong magnetic fields dynamic is observed, generated itself by these movements. In some peculiar cases, the magnetic field will take the shape of a tube and block the convection, forming spots surrounded by faculae (Schrijver & Zwaan 2000; Karttunen et al. 2007). From solar observations, these latters are more extended by a factor $\sim$ 10 in comparison to spots (Chapman et al. 2001). Because only a small amount of the stars' surface is covered by these activity's evidences, only a small fraction of the total flux is affected, producing a RV signal largely below the km/s. As an example, at maximum, the Sun's surface is covered at 0.1% by spots and 1% by faculae producing roughly a RV jitter of a few m/s (Borgniet et al. 2015).

Several activity features are visible in Sun's observations, the main are drawn in the sketch Fig.4. We separate the solar atmosphere in three zones : the corona, the chromosphere and the photosphere. For stellar RV contamination, only the latter two are relevant. The chromosphere is the region were plages are found, formed by the summit of magnetic field loops. However, this layer is not emitting a lot of flux and only a few lines are formed in this region due to the very high temperature. The photosphere, as suggered by its name, is the region where most of the photons are coming from. The majority of the activity features contaminating RV are dwelling there like spots, faculae and granulation (Fig.5). Often, plages and spots are observed to appear by pairs of opposite polarity because they are formed by the same magnetic loop.

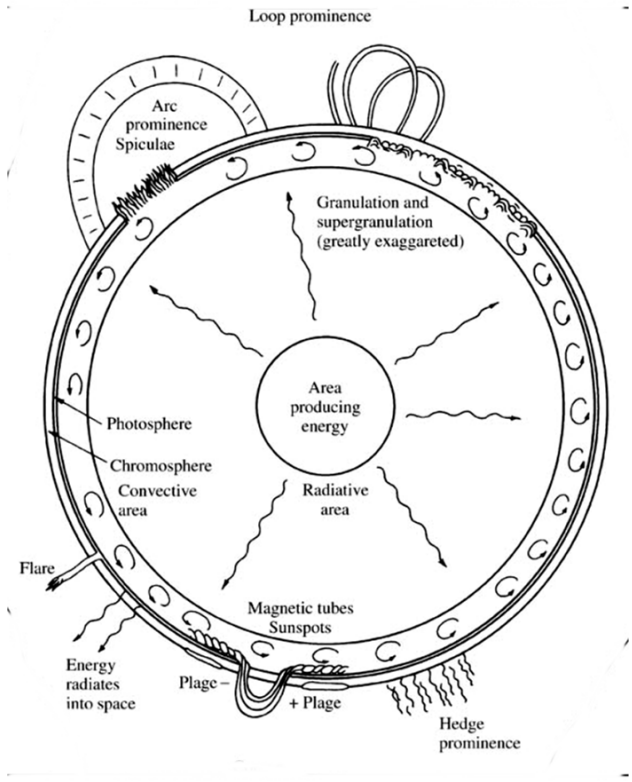

Fig. 4: Sketch of the main components of solar activity. The Sun is made of a radiative core surrounded by a convective envelop. Two thin regions are to point out : the photosphere and chromosphere. The latter one contains plages, whereeas the former contains spots, faculae and granulation. Spots and plages appeared often by pair of opposite polarity linked by the same magnetic loop. (*Credits : Van Zandt, R.P. (1977):Astronomy for the Amateur, Planetary Astronomy, Vol. 1, 3rd ed.*)

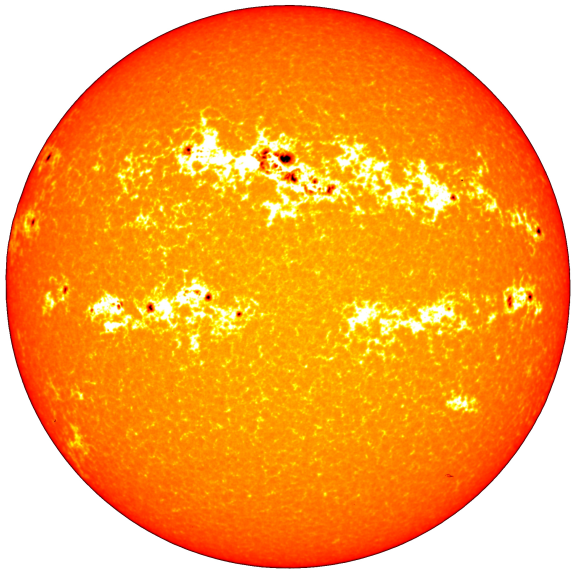

Fig. 5: Visible observation of the Sun's photosphere while the maximum of the 23^{th} Schwabe's cycle, at the 28 March 2001. The granulation gives the patchy aspect to its surface whereas sunspots (*dark spots*) and faculae (*bright regions*) are linked with strong magnetic fields evolving at its surface. (*Credits : NASA/Goddard Space Flight Center Scientific Visualization Studio. Source data courtesy of HAO & NSO PSPT project team*)

We distinguish four types of stellar contamination in RV measurements (Dumusque 2012) : 1) the granulation and supergranulation (already discussed) 2) the short acoustic oscillations 3) the cold spots and hot faculae 4) the long magnetic cycles. The stellar oscillations (sometimes called «star's breathing») contribute the less and can be reduced taking exposure longer than the typical period of variation ($\sim$ 5 min for the Sun) or averaging several observations over one night. The long trend variations are perturbing the detection of long period planets on Sun-like stars with variations over several years, whereas on the timescale of dozens of days, spots and faculae induce RV variations at a level of several m/s.

An active region affects the RV in two distinct ways (see Dumusque et al. (2014) for more details). The first perturbation is linked with the flux modification coming from the active regions. A spot has a smaller temperature than the surrounding surface, as a consequence the flux emitted from this region is smaller and the balance between the right and left part of the star is perturbed. From the phenomenological aspect, the effect is similar to the Rossiter-McLaughin effect of a planet in transit. Let us imagine a situation as displayed on the left of the Fig.6. For an immaculate surface, for each region on the left there is a region on the right compensating the rotation of the star (lines are broader but not shifted). When a spot appears, the balance in flux is broken such that a small amount of blueshifted flux is missing, producing an apparent redshift, implying an increase of the RV. When the spot reaches the right part the effect is reversed. The reasoning is inverted for faculae since their temperature are higher and thus more flux is emitted.

The second modification is the inhibition of the convection by the strong magnetic fields (Meunier et al. 2010). In standard condition, the convection provides a blueshift contribution since major part of the convection cells is moving toward us (right part Fig.6).

Finally, depending on the rotational velocity and on the stellar inclination either the flux effect or the convective blueshift inhibition will dominate. The RV contribution of spots is linked with the projected stellar rotation $v_{rot} \sin i$, whereas the convective contribution of faculae is related to the velocity matter upflows $v_{z,flow}$. Thus, the main RV contamination for slow rotating stars or pole-on stars comes from faculae. On the opposite, spots are dominating the RV contamination for fast rotators observed equator-on. A limit of $v_{rot} \sin i = 8$ km/s can be fixed to separate equator-on stars that are faculae or spots-dominated (Dumusque et al. 2014).

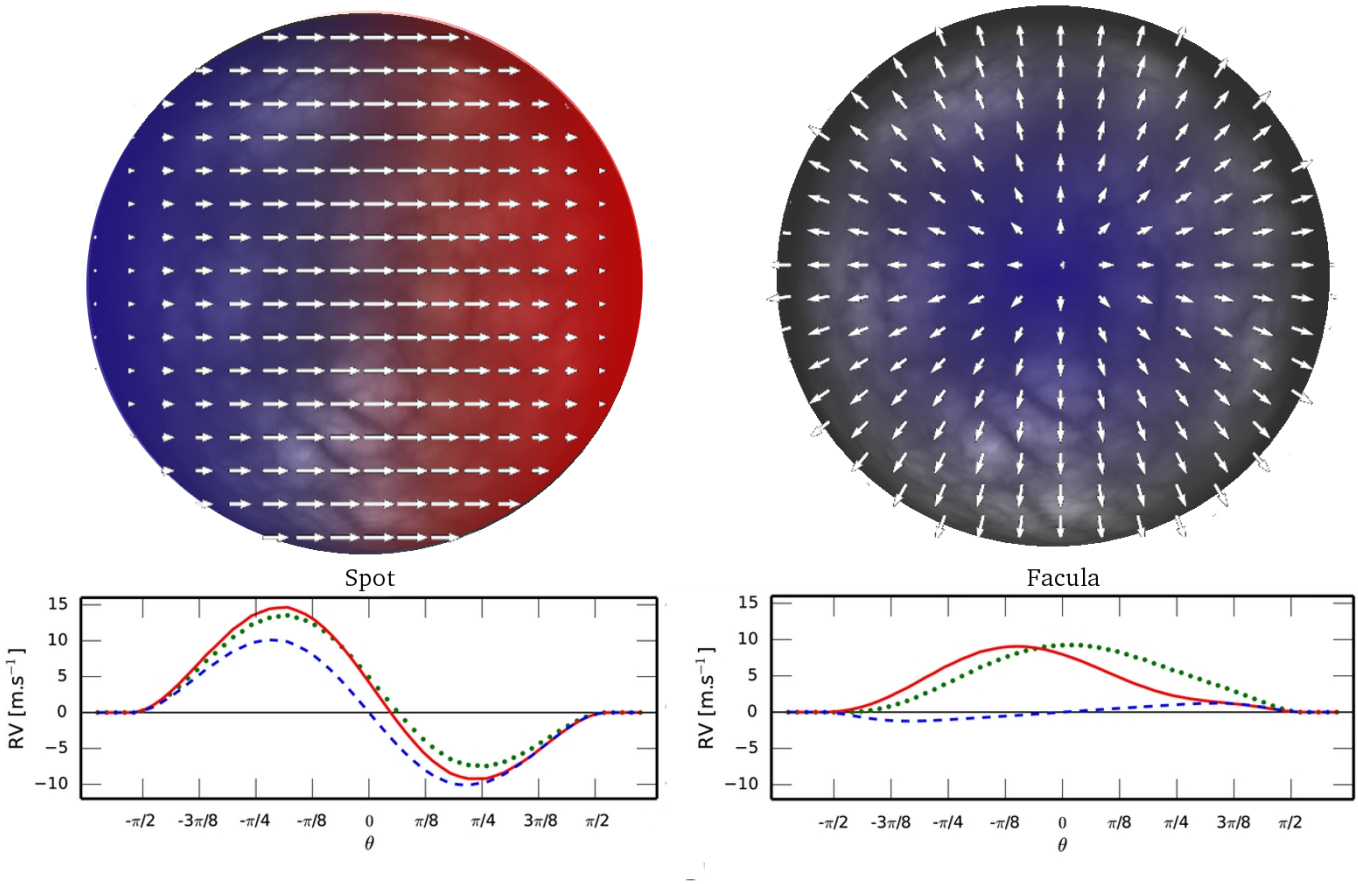

Fig. 6: Schematic effect of an equatorial spot (left) or facula (right) with a filling factor of 1% on the Sun (projected rotational velocity of 2 km/s and star seen equator on). For the simplicity of the sketch, we neglect the horizontal component of the convective velocity and we consider a purely vertical velocity. The vector field indicates the tangential component in the sky of the velocity vector (not affecting RV) whereas the radial component is displayed by the color code. **Left** : Effect induced by the rotation of the star. **Right** : Effect induced by the motion of convection (convective blueshift). **Bottom** (*Crédits :* Dumusque et al. (2014)) : Simulations by the SOAP 2.0 code of the RVs as a function of the limb angle θ ($\theta = 0$ at center and $\theta = \pm\pi/2$ on the limb) induced by a spot (left) and a facula (right). Three models are displayed : neglecting the convective blueshift (*blue dashed curve*), with a convective blueshift similar to our sketch and independent of the depth (*green dotted curve*), lastly a more complex model based on solar observations (*red solid curve*). For spots, the effect dominating the RV is the perturbation of the flux balance between the red and left part of the star. For faculae, the main contribution comes from the inhibition of the convective blueshift. But both effects are actually contributing which explains, for instance, why the RV signal induced by the spot is not perfectly null at $\theta = 0$ for a realistic model.

3. Individual radial velocities

3.1. The standard method (CCF)

State-of-art of radial velocity measurements have exploited the properties that the mutual gravitational attraction between star-star or star-planet should affect all the lines with the same drift (more precisely by the same lag as explained below) given by the Doppler shift formula in the classical regime :

$$\frac{\lambda - \lambda_0}{\lambda_0} = \frac{\Delta\lambda}{\lambda_0} = \frac{v}{c} \tag{1}$$

A question is to know precisely how to measure this shift. A spectrograph, in its more simple description, is composed of a prism to split the different spectral components of the light and of a detector to measure the intensity of the components. The classical detector used is a CCD. In the absolute, a CCD does not provide any informations on the shift since a CCD is a measurement of intensity on a pixel grid. To get the shift, the CCD has to be first calibrated such that after the calibration a pixel versus wavelength relation is established (so called the wavelength solution). For instance, a Thorium lamp is used in this purpose for spectrograph in the visible. Thorium produces lines at well known locations by atomic physics (because energy level are known and thus the wavelength transition also) which allows to determine the correspondence between the position on the detector and the wavelength. This calibration can be extended on the full detector (even if a region is deprived of Thorium's lines) by fitting a polynomial function which is similar to assume that the prism's dispersion is smooth.

To average out the RV signal from all the lines and therefore increase the precision on the stellar velocity, the current state-of-art is to perform a cross correlation between a mask $t(\lambda)$ and the stellar spectrum $s(\lambda)$. This mask is 1 at wavelength were there is no lines and 0 at the lines' center. The cross correlation function (CCF) being the convolution between the mask and the spectrum :

$$CCF(\Delta x) = (s * t)(\Delta x) = \int_{-\infty}^{\infty} s(x)t(x + \Delta x)dx$$

Let us precise that because we are dealing with a grid (the pixels), the convolution can be performed in its discrete writing. Also, the CCF can be optimised considering that more RV informations are given by sharp deep lines which introduce weighting coefficients (Pepe et al. 2002). The reader may remark that we did not write explicitly $\Delta\lambda$ but Δx (called the lag) since, for a fixed radial velocity, the shift $\Delta\lambda$ depends also on the wavelength λ_0. To solve it, the scale used for the CCF is generally logarithmic such that $\Delta x = ln\lambda - ln\lambda_0 \equiv \Delta ln\lambda$.

Translating in common words, the CCF measures how the overlap between the spectrum and the template change in function of the template's lag. When the template's shift equals the Doppler shift, if the template has been well formed, the spectrum and the template are perfectly superposed and so the value of the cross correlation is at minimum. The RV is then given by the barycenter of the CCF, which is generally measured by fitting and inverse Gaussian to the CCF and then taking its mean. This method is particularly efficient since the RV is measured using all the lines' information at the same time (often we speak about the CCF as a kind of average line).

What if each spectral line is affected differently by stellar activity, and therefore have different behaviours ? Indeed, since each line is formed at a different optical depth level, has a different sensitivity to temperature variations, they should have a different behaviour in the presence of spots and faculae due to different temperature and magnetic field strength that inhibit locally the convection. Each line will therefore have its own signature $RV_{a,i}$ that depends on the properties of active regions (Davis et al. 2017). On the contrary, a planet must induce exactly the same RV signal RV_p on all the lines because the shift in wavelength is caused by an external process not related with the line formation : the mutual gravitational force between the star and the planet. We can thus decompose the RV of a line (RV_i) as the superposition of the planetary and stellar RV signal :

$$RV_i = RV_p + RV_{a,i} \tag{2}$$

The consequence of the second term is to modify the CCF in a sensitive way, either shifted, larger or even introducing an asymmetry. At this point, the CCF begins to suffer of its previous advantages and we may thought to restrain the lines used in the mask. But for that, we need to know the individual radial velocity of the lines which is not an observable generally derived.

3.2. New method

Let us take an analogy to illustrate the idea of our work (Fig.7). Let us assume that a cannon as been installed in a field and is pointing fixedly toward an empty position with a known angle. Our goal is to estimate the velocity of the projectiles expulsed by the cannon using only their impact location (a classical ballistic problem). The trajectories are not solely determined by the cannon but can be affected by the recoil. This perturbation will at the end spread the impact points such that using only one final position, our prediction is really bad on the initial velocity. However, shooting a lot of cannonballs, the initial velocity can be recovered using the statistical average impact point of the distribution. In this metaphor, the initial velocity of the shoot is the planetary signal and the bullets are the lines (their final impact being their RV), whereas the recoil ensures a kind of "noise". In definitive, even if the RV of the lines reach several meters or even hundred meters per second, when a lot of lines are used their mutual information can constrain efficiently the RV. Moreover, each line possesses an uncertainty σ (let us say represented by the size of the hole in the ground). When the average will be taken, it is a reasonable choice to give more "weight" on the most precise lines (smallest holes) compared to the other.

Suddenly, a gust of wind is raising from our back and unfortunately we remark only now that all the balls are actually not identical ! Some balls are heavier than the others such that the lightest one are more deviated by the winds (we remind that the maximum distance is independent of the mass in windless condition). A light ball in this story is the equivalent of a line strongly affected by activity. As a consequence, instead of the previous pattern of position, the wind is pushing a bit farther the lightest ball such that after computation the initial velocity is found to be higher than the true one. This is precisely the kind of perturbations induced by stellar activity. The idea is thus the following : if there is a way to know which balls are heavier, we could either neglect the lightest one to focus only on the impact of the heaviest or correct the position of the lighter. For the first idea, the price to pay is to lose a bit of precision because less balls are used, but as we will see the gains fill losses. That is the main idea of this work, find a way to distinguish lines affected and those less affected by activity such that at the end the contamination from stellar activity can be reduced from the RV. To be able to do this, we need to measure the RV of every spectral line individually.

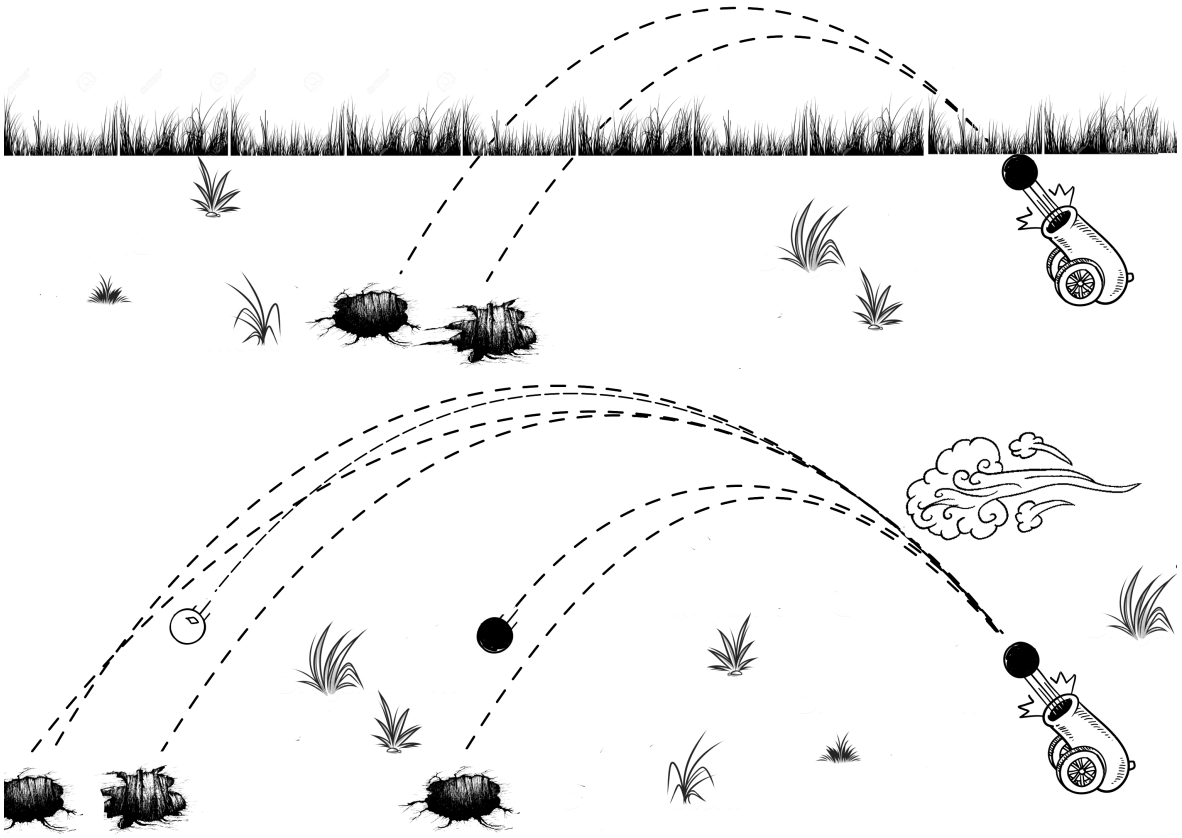

Fig. 7: Metaphor illustrating the main lines of our work. **Top** : The RV velocity of the planet (*initial speed*) as well as the noise (*recoil of the cannon*) are both contributing in the RV of a line (*maximum distance of a cannonball*). The precision using one ball is small to compute the initial speed, nevertheless, combining the measurement of hundreds or thousands of shoots, the maximum distance can be determined more precisely through a weighted average, where the weight is inversely proportional to the uncertainty (*the hole's size and its irregularities*). **Bottom** : The stellar activity (*the wind*) perturbed the previous picture affecting strongly some light balls (*white ball*) and less the heaviest (*black ball*). To recover the planetary signal, we need to identify the lines less affected by activity.

We used the reduction given by Dumusque et al. (in preparation), which provides the RV of each spectral line (RV_i) present in the K5 template as well as their uncertainty. We explain briefly the different steps to compute the average radial velocity of a choosen selection of lines. If all the lines are used, we will explicitly use the RV_m notation which corresponds to the result of the CCF. We bin the RV_i by night of observations thanks to a weighted average. Note that in the analysis that we will perform, the stars have rotational periods of dozens of days, therefore, binning the data allows to increase the precision, mitigate short-term stellar signals without affecting activity signals. The individual weights w_i are computed from the squared inverse of the photon noise error bars σ_i on the RV_i :

$$RV_{i,binned} = \frac{\sum w_i \cdot RV_i}{\sum w_i} \quad w_{i,binned} = \sum w_i$$

Once the RV_i are binned per day, we can then measure the RV from all the lines, RV_m, by performing a weighted average. Finally, a first or second polynomial fit can be performed to remove potential trends not linked with the star activity but rather by magnetic cycle or companion. Once in possession of the RV_m (top right subplot Fig.8), we compute the correlation between the RV of each lines and this signal. To measure the correlation between the RV_i and RV_m, the RV_m's trend is removed from the binned RV_i (bottom left subplot Fig.8). Eventually, a weighted linear correlation is performed between the RV_i and the RV_m (bottom right subplot Fig.8). The RV_i are put on the y-axis since their errors are bigger than the average ones. Two quantities are measured : the Pearson coefficient $\mathcal{R}$ and the slope of the regression. The Pearson coefficient measures how close to the average signal is the signal measured on RV_i. If the signals are perfectly similar in shape, the slope quantifies the scaling factor to apply on the RV_m to reach the RV_i.

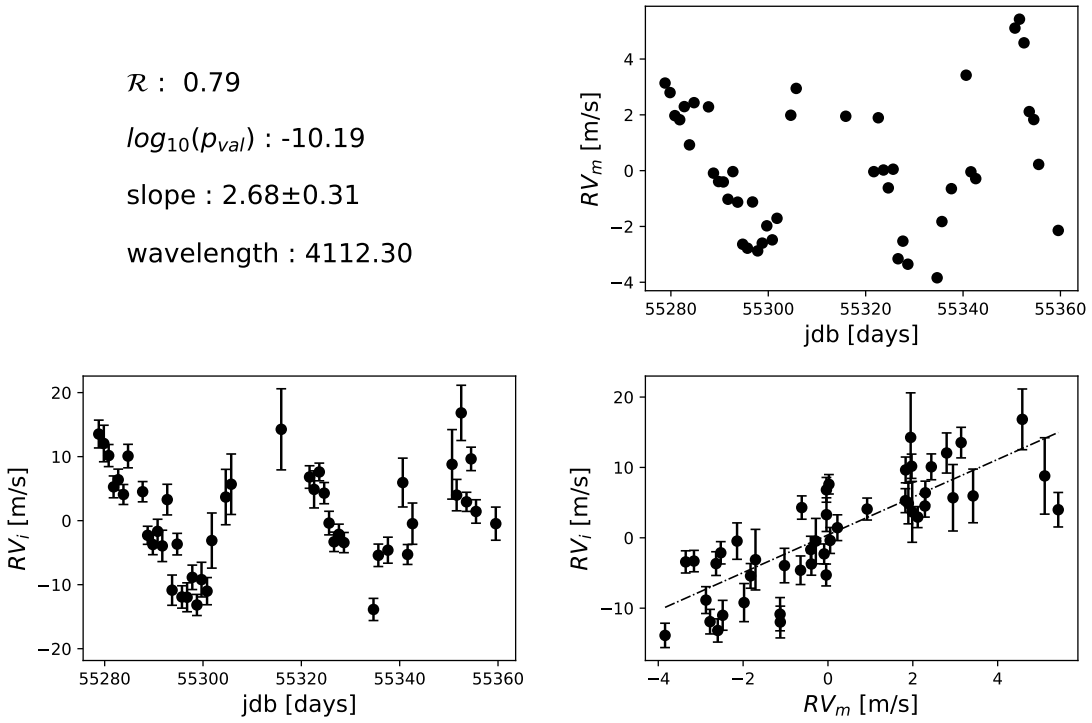

Fig. 8: **Top right panel** : Detrended 1-day binned average radial velocity RV_m of α Cen B measured on all the spectral lines. **Bottom left panel** : Detrended 1-day binned individual radial velocity RV_i of the line at 4112.30 Å. **Bottom right panel** : Weighted linear regression (*dashed dotted line*) between RV_i and RV_m. The coefficient of correlation $\mathcal{R}$ as well as the log p-value, the slope and the wavelength from the HARPS K5 mask (round at two decimals) are indicated.

4. Classification based on flux variation

4.1. Overview

We followed the analysis presented in Thompson et al. (2017). The authors report to have observed flux variation inside the lines. An interesting observation is that these temporal variations appear correlated with the stellar activity itself. Hence, a legitimate question is to investigate if the lines showing significant flux variation are also the lines inducing the most important RV variation. Going further, is the shape of the flux variation playing a key role or not ?

The target was the first nearest and third apparent brightest star system α Centaurus, more precisely the K1V star HD121628 (abbreviated as α Cen B), whose its brightness ($m_V \simeq 1.3$, Karttunen et al. (2007)) makes it as a perfect candidate for precise RV measurements. Indeed, a large amount of light is needed when we want to spread the light for high resolution spectrograph without being dominated by photon noise. HD121628 is a slightly cooler star ($\sim 5200°$ K, Sousa et al. (2008)) compared to the Sun and exhibits hints of solar-like activity at its surface as expected for cooler stars.

Instead of working directly with the lines themselves, Thompson worked with ratio spectra what could be a priori dangerous because the information is now coming from two spectra instead of one, and the two informations can be mixed in a complex way. But if the spectrum use to divide is kept fixed, it acts more like a normalisation process and reduces the mixing of the informations. One advantage of this method is to highlight variations between high and low activity level, no matter the initial profile of the line. In counterpart, the information on the initial line profile (strong, weak, saturated, assymetric) is lost and the noise increased.

We analysed 1767 s1d spectra from the HARPS spectrograph ($R \sim 115'000$) taken on 47 different nights in the visual domain between 3781 Å and 6913 Å with a sampling of 0.01 Å. The 1767 spectra cover roughly 2 periods of rotation during the year 2010, which represents ~ 80 days. The rotational period has been deduced from the $\log R'_{HK}$ index based on the equivalent width (EW) of the emission seen in the core of the Ca II H and K lines (3933 and 3968 Å lines) known to be correlated with the activity (Noyes et al. 1984). Such emission lines are produced during chromospheric heating events. The activity index shows a clear and periodic modulation thought to be induced by the presence of an important active region rotating at the surface of the star. As generally done (Wright et al. 2004), the rotational period is recovered using the $\log R'_{HK}$. To find out the rotational period of α Cen B at the latitude of the active feature, a simple periodic sinus with a linear drift signal has been fitted. The Monte Carlo Markov Chain (MCMC) which has been executed on the time variation of the activity index (Fig.9) provides a period P_{rot} of 37 days. According to Dumusque (2014), the activity signal is compatible with a huge facula rotating at a latitude of 44° for a size covering 2.4% of the surface, even though some degeneracy exist between the latitude and the inclination. From the observed signal, it seems that the facula is always visible. This can be explained by the fact that α Cen B is seen with an inclination angle of 22° which mean almost pole-on (see Fig.10).

To work with the full sample of spectra, we must first cut them such that every spectrum begin and end at the same wavelength. Indeed, even though the grid of wavelength is the same for all spectra, each spectrum begins and ends at different wavelengths. We take as first wavelength the maximum of all the minima ($\lambda_{min} = 3781.77$ Å) and we cut at the end with the minimum value of all the maxima ($\lambda_{max} = 6912.69$ Å).

We began by reproducing the Fig.1 of Thompson et al. (2017). All spectra are cut between 4370 and 4410 Å to get and overall view of the different behaviours observed. The first step is to form a *master quiet spectrum* (MQS) grouping nights with the lowest activity index. The MQS represents the average or the « typical » spectrum of α Cen B when the star is not active (assuming that the spectra are not strongly variating). In our case, we took the 10 percentile spectra with the smallest $\log R'_{HK}$ value, which corresponds roughly to the 7 nights around BJD=55294 on Fig.9. Each spectrum is first normalised in flux, since all the spectra have not the same absolute flux. At the end, the average spectrum (MQS) is computed taking the median between the normalised spectra.

We then chose 13 roughly equidistant nights that probe the activity during its different phases, these observational nights were the 1, 6, 11, 17, 23, 25, 26, 30, 34, 38, 40, 41 and 46^{th}. The same analysis is reproduced : normalise each spectrum in flux before to take the median over the nights. At the end, we divided each of this "average" spectrum by the *master quiet spectrum* and plot the results in Fig.11. On this figure, we clearly distinguish behaviours which look by eye correlated with the activity level, such as the narrow "pseudo-emission" peaks. The term "pseudo-emission" emphasises the fact that because we are dealing with relative spectra, peaks are not necessarily linked with a physical emission line, but can also highlight a line becoming weaker when the star is active compared with its quiet phase.

To illustrate the diversity of variation, we show the 4375 Å pseudo-emission peak in Fig.12 as well as the 4616 Å pseudo-absorption peak in Fig.13. We showed that the ratio profile is modulated by the activity computing the «pseudo - EW» (again we use the word pseudo to clearly indicate that we are working with the ratio profile). The center of the line was provided by a theoretical mask (HARPS K5) and the area computed on a width of 0.1Å with trapezoidal rule integration. The correlations between the $\log R'_{HK}$ coefficient and the pseudo - EW were excellent ($\mathcal{R}_{pearson} = 0.98$ and $\mathcal{R}_{pearson} = -0.90$) and matching with the observation of Thompson et al. (2017).

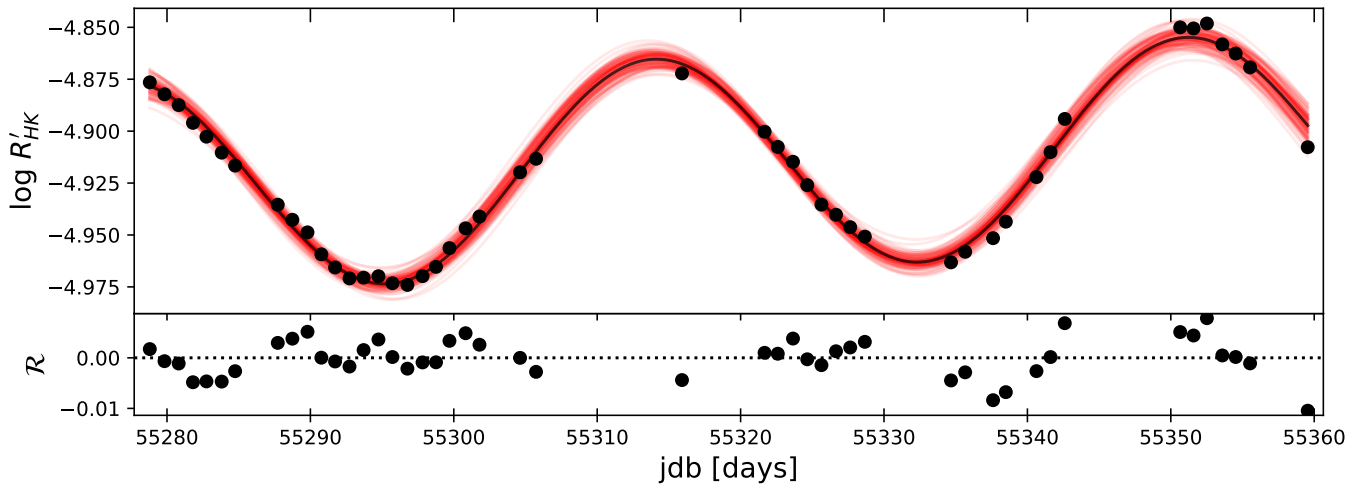

Fig. 9: Determination of the rotational period of α Centaurus b ($P_{rot} = 37.2 \pm 0.5$ days) with MCMC using the log log R'_{HK} activity index as proxy. The model used is a sinus with a linear drift. **Top panel** : Time variation of the activity index (*black dots*). Until now and for the end of this paper the days are in jdb - 2'400'000. The model using the median of the parameters fitted is shown (*black solid line*) whereas one hundred of random individual simulation are drawn (*red shadowed envelop*). **Bottom panel** : Residual from the median of the fitting.

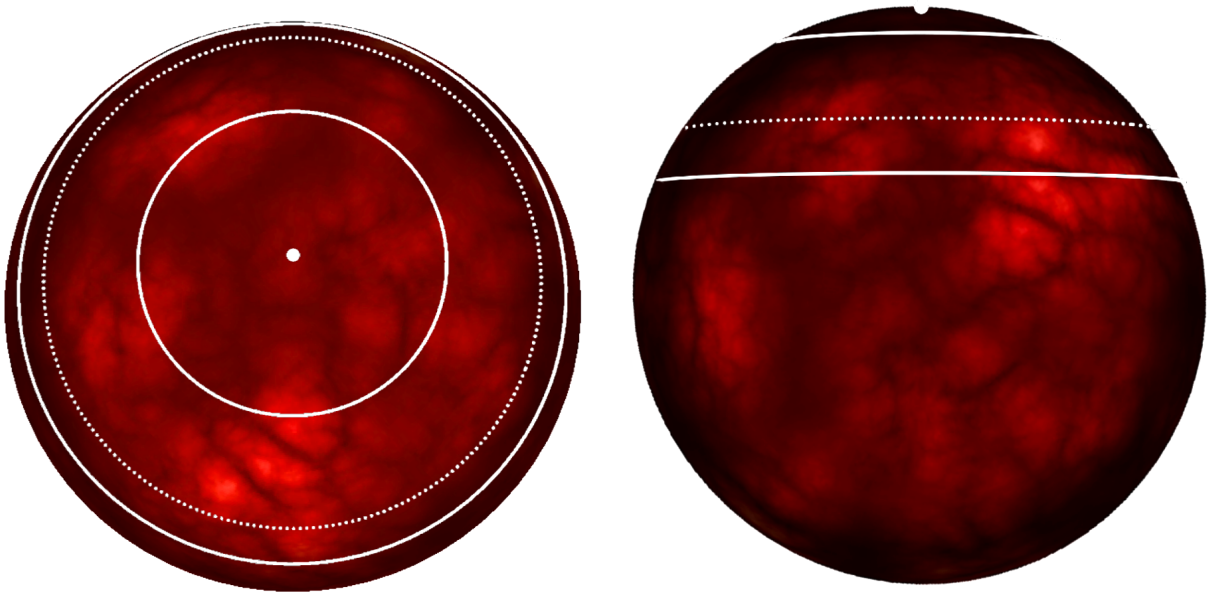

Fig. 10: Artist view of the plage's position at the surface of α Cen B given by the best fitted model of Dumusque et al. (2014). The pole (*white dot*) is almost aligned with the Earth such that the plage at the 44° latitude never disappears (*dashed white line*). For visualisation, two others parallels are displayed at 22° and 60° (*white solid lines*). Let us precise that, in the model used to predict the configuration, some degeneracy exists between the latitude of the regions and the inclination angle. **Left** : View from the Earth. **Right** : Equatorial view.

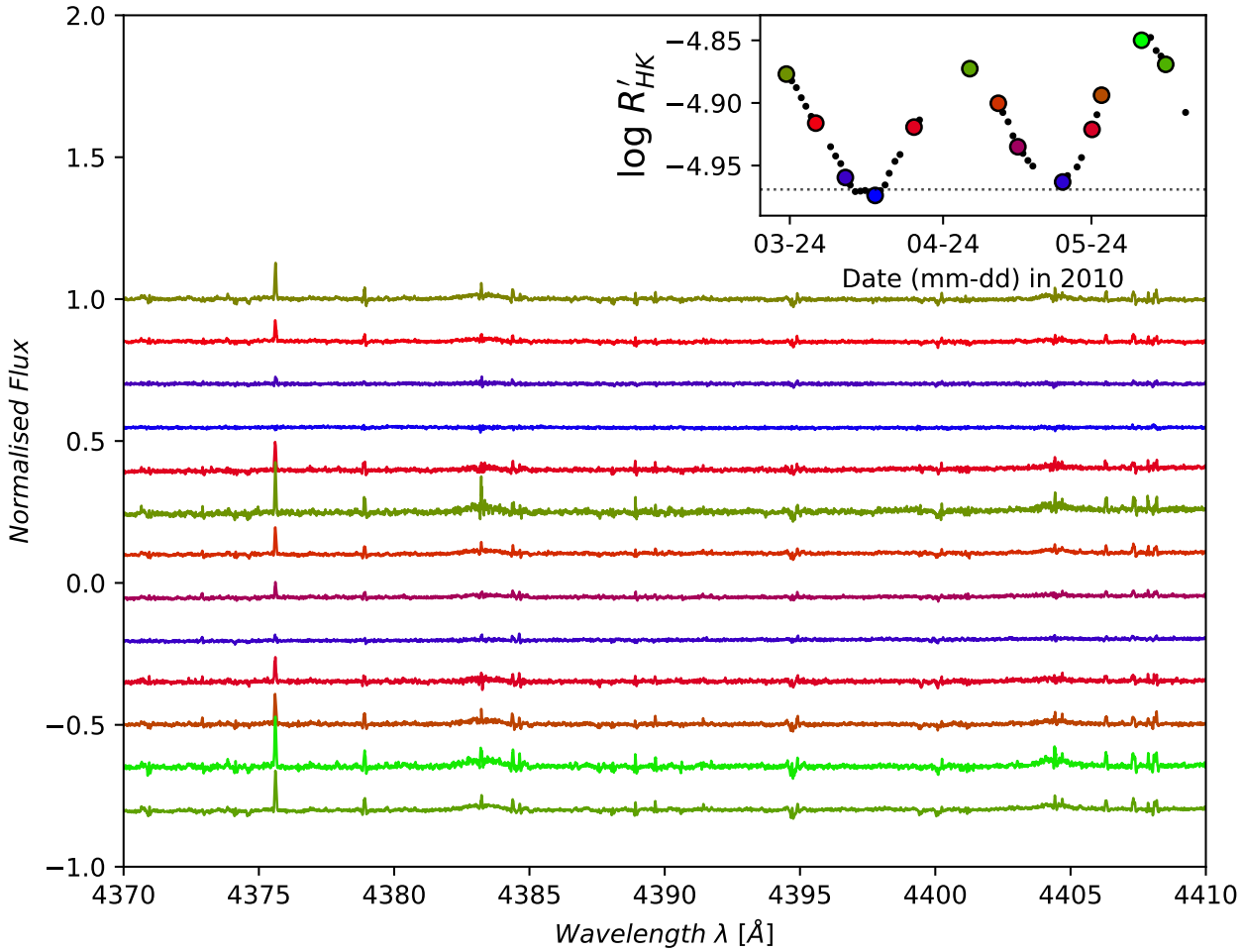

Fig. 11: Reproduction of Fig.1 from Thompson et al. (2017). A clear focus on the variation of the line profile in our dataset of ratio spectrum. **Inner plot** : variations of the activity index as a function of time on the 47 observational nights (*black dots*). 13 roughly equidistant nights are chosen to probe the different phases of activity (*color dots*). The color code goes from blue (quiet) to green (active). The 10 percentile limit to choose the data selected to build the master quiet spectrum has been drawn (*dotted line*). **Outer plot** : Ratio spectra (median spectra over night normalised by the master quiet spectrum) for different level of activity. The color code is the same that the one described in the inner plot. An arbitrary offset has been added for each different ratio spectrum for clarity. Several features can be observed which seems by eye modulated with the activity level of the star. The strong narrow emission peak at 4375 Å, as well as the broad emission at 4383 Å are both attributed to FeI lines according to Thompson et al. (2017).

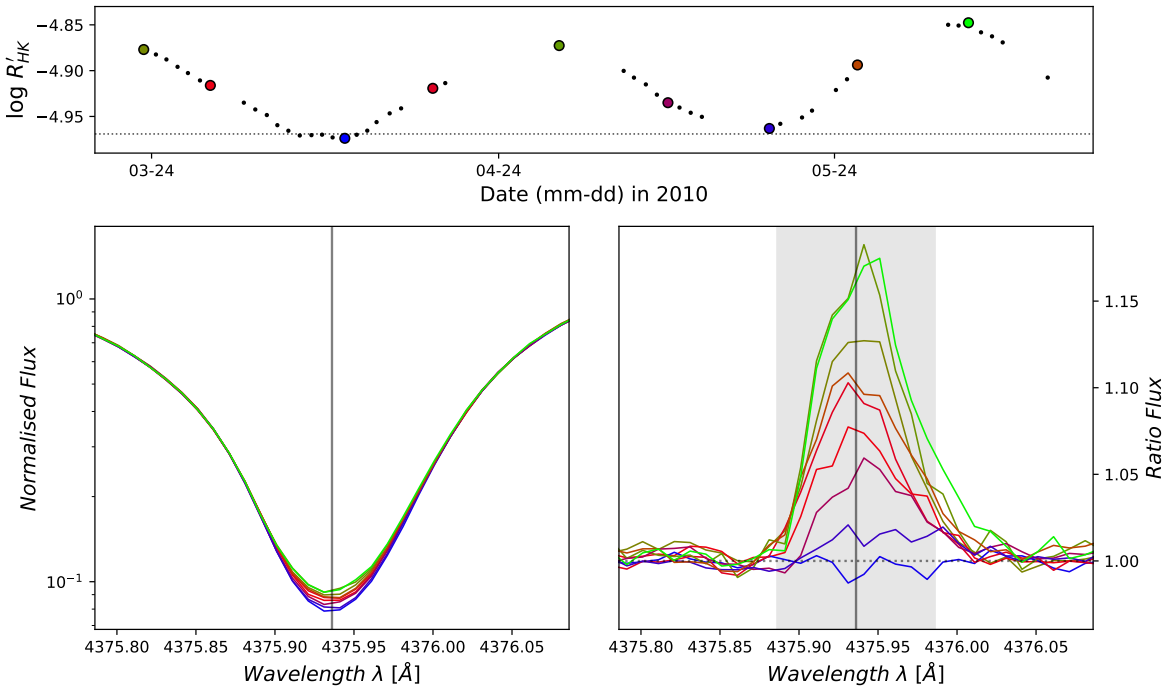

Fig. 12: Closer look on the very deep 4375 Å FeI line. **Top panel** : time variation of the activity level (***black*** *dots*), the 10^{th} percentile level is indicated (*dotted line*). **Bottom-left** : Flux normalised of the strong absorption FeI line. The line clearly becomes weaker when the star is more active. **Bottom-right** : Closer look on the "pseudo - emission" peak in the ratio spectrum which shows a variation between high and low activity level of 17%. The area where the "pseudo EW" where computed for the further computations is displayed (*shadowed area*). This pseudo - EW is clearly correlated with the activity index ($\mathcal{R}_{pearson} = 0.98$). The center of the line, given by the HARPS K5 catalogue (shifted from the radial velocity of the star system : $V_{rad}^{system} = -22.7$ km/s) , is situated in the middle of the area (*vertical gray line*).

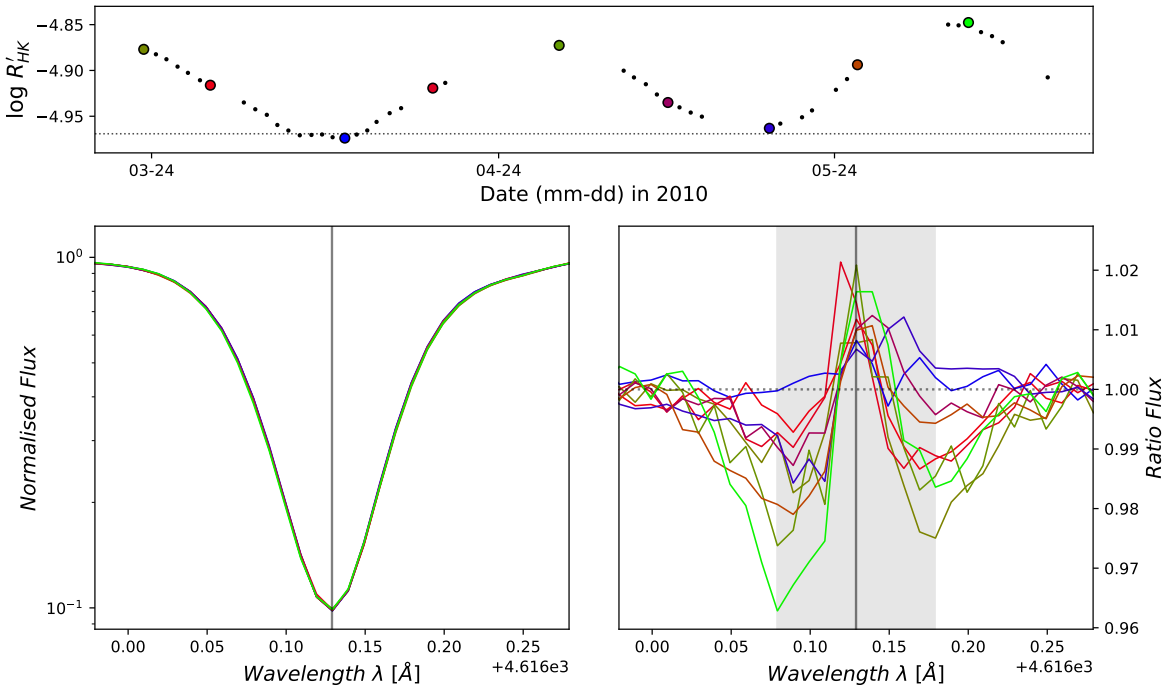

Fig. 13: Same as Fig.12 for a 4616 Å CrI line exhibiting a broadening modulated by the activity. The pseudo EW is found anticorrelated ($\mathcal{R}_{pearson} = -0.90$) with the activity level since the EW is signed.

4.2. Significant line variations

We reproduced the work of Thompson on the full wavelength domain. Each spectrum is first shifted in wavelength (the shift depends on the time of the observation) to cancel the RV linear trend effect of the companion α Cen A (see Fig.19). Because after the shift each spectrum possessed its own «shifted grid», we returned to the initial grid (which was the same for all the spectra) using a cubic interpolation. The whole spectral range in our possession has been divided in 95 segments of 36 Å, an overlap between them was computed to match the first and last segment (border left and border right respectively) with our minimum λ_{min} and maximum λ_{max} wavelength. We then investigated each segment independently, their flux have been normalised.

We created as before the same *master quiet spectrum* (MQS) by grouping the 10 percentile spectra with the smallest $\log R'_{HK}$ index, we created also a *master active spectrum* (MAS) by taking the 10 percentile of highest $\log R'_{HK}$ index. It covers 6 nights around the 55353 BJD. In hindsight this choice to take active spectra on several nights is largely questionable. Indeed, for the non active period there is a priori no deep reasons for a spectrum to change and so to find large differences between spectra inside the sample forming the MQS. However, during active period, spectrum might be modified in a sensitive way such that the information can be spread when spectra are averaged. We concede that the following assumption claiming that during six nights of activity the spectra are sufficiently similar to be average can be litigious and we keep in mind this aspect. On the opposite, if a weak peak or asymmetric peak is founded, it won't be possible to attribute them to noise or unlikely events, because these features were present on several days, giving them more "weight". We precise that we investigated the impact of this choice in the appendix B which appeared marginal.

We computed the ratio spectrum between the MAS and the MQS (the signal obtained is thus similar to the light green one in Fig.11). The significance of features is established thanks to a moving median and a moving standard deviation using a rectangular smoothing kernel, which consists in placing a box (width of 3Å) centered at a fixed wavelength in which the median and the standard deviation of the ratio flux are computed. The box then "slides" along the wavelength axis computing these two values for each new position. A peak is defined as every local maximum ($flux(\lambda_{i-1})<flux(\lambda_i)>flux(\lambda_{i+1})$), whereas a peak is significant if it "emerges" outside a 3σ threshold around the median. The standard deviation has been computed with the 16^{th} and the 84^{th} percentile. The lines are then identified using a theoretical mask K5 (similar type star) shifted by the radial velocity of the α Centauri system derived by Wesselink (1953) ($v_{rad} \simeq -22.7$ km/s). The results of these operations are drawn, for the same spectral range than the Fig.11, in Fig.14.

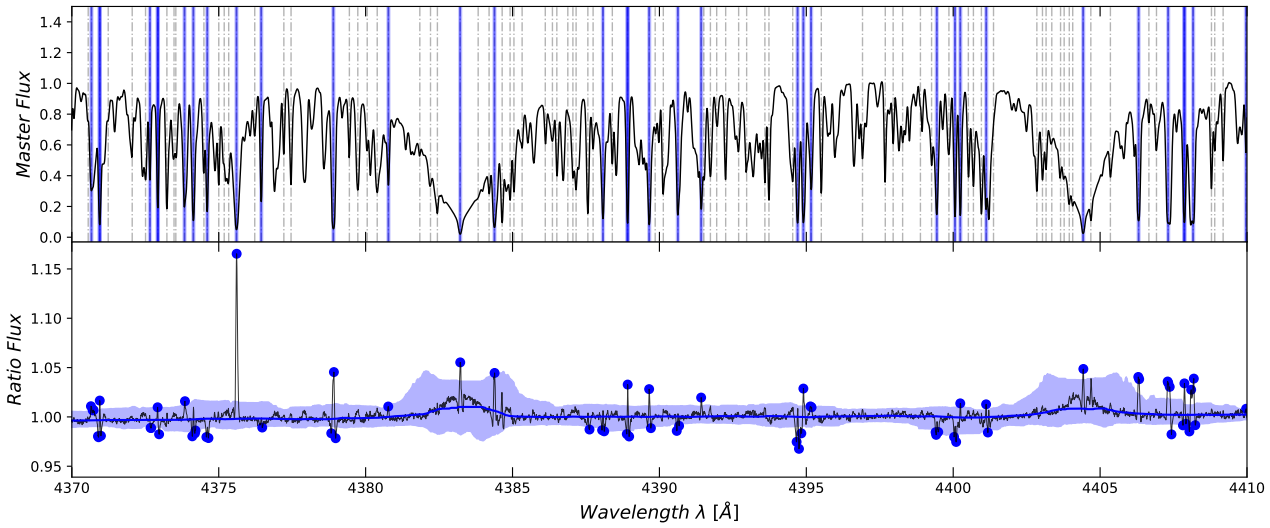

Fig. 14: **Top panel** : Master quiet spectrum normalised. A catalogue of lines K5 is displayed to identify the lines (*vertical dashed gray lines*), those which are at a distance smaller than 0.05 Å of a significant peak (*blue points*) are considered as significantly affected (*blue lines*). **Bottom panel** : Significance criterion. The ratio spectrum MAS/MQS (*solid black line*) exhibits peaks indicating variation in theit line profile. A moving median (*solid blue line*) and a moving standard deviation were computed along the wavelength axis. If the ratio spectrum goes outside the moving 3σ threshold (*blue envelop*) the peak is considered as significant (*blue points*).

The envelop's width (standard deviation) is linked with the width of the box used and therefore more peaks are detected if the box is larger. The kind of peaks selected is also directly linked with the box. In our case, the broad features are not detected by our criterion, but the narrow ones are. Nevertheless, we can remark that broad features could be detected visually as a spread of the envelop. As we want to focus on the narrow ones, this 3Å box sounds as a good choice. We defined 6 groups that we explain below :

1. no variation
2. positive variation
3. negative variation
4. both (thunder variation)
5. positive asymmetric variation
6. negative asymmetric variation

If a line of the K5 catalogue has no detection peak in the spectrum ratio, it means that the line does not vary significantly with time and therefore with activity, hence it is flagged as a « *no variation* » line. All the gray dashed lines on the top panel of Fig.14 are of this kind. The other classes are based on two quantities : the height of the ratio peak or the pseudo EW.

The first class is based on the height of the peak and provides the Υ parameter. Because we flagged each local maximum, we first regrouped all the peak attributed to a line from the HARPS K5 line list. As said previously, if there is no significant peak, the line is in the « *no variation* » group. When there is only one peak in the vicinity of the line, the line is either « *positive* » (a pseudo emission peak, $\Upsilon \equiv 1$) or « *negative* » (a pseudo absorption peak, $\Upsilon \equiv -1$). If two peaks or more are present, the task is more difficult. When all the peaks are in the same direction, we just kept the highest/lowest one, giving a positive/negative line. If positive and negative peaks are both present, more steps are needed. Let us take an example, assuming there are 3 positive peaks and 2 negative ones flagged around a given line. We keep the highest p_+ and lowest p_- peaks values (more precisely the distance between the height and the moving envelop) before to compute a kind of «contrast», defining a first asymmetric parameter Υ_{asym} :

$$\Upsilon_{asym} = \frac{|p_+| - |p_-|}{|p_+| + |p_-|} \tag{3}$$

The advantage of this parameter is that when $p_+ \gg p_-$, $\Upsilon_{asym} \longrightarrow 1$ what is similar to a positive line. On the opposite, if $p_+ \ll p_-$, $\Upsilon_{asym} \longrightarrow -1$, thus a negative line. If the ratio peak profile is perfectly antisymmetric ($p_+ = p_-$), $\Upsilon_{asym} = 0$. We display more explicitly how this parameter works in Fig.15, by attributing a color code to the Υ parameter. The code goes from *negative* (blue) to *positive* (red), passing through the negative/positive (thunder) shape in green (i.e lines that are $+/-$ or $-/+$).

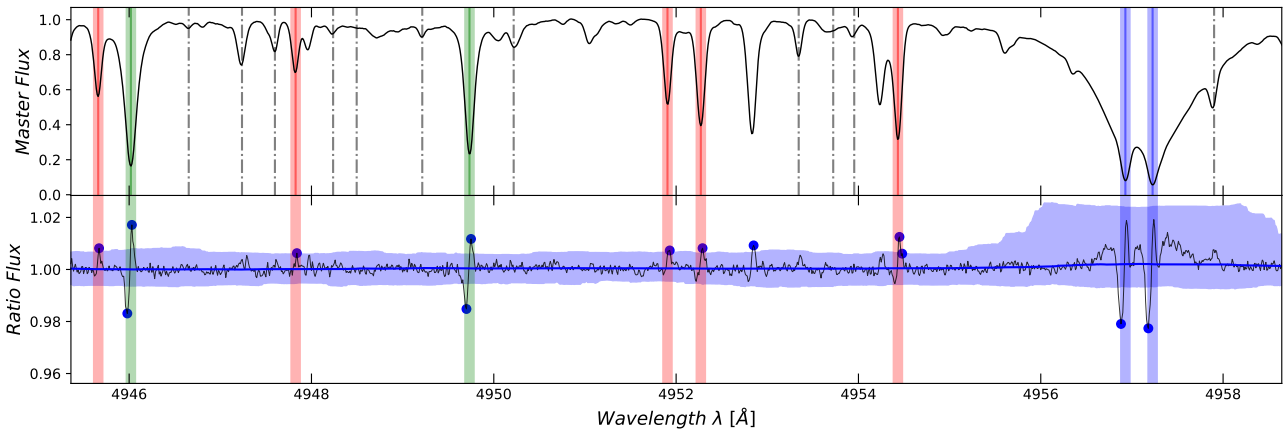

Fig. 15: Same figure than Fig.14 on another spectral range adding the results of the Υ parameter attribution. A *positive* line ($\Upsilon = 1$) is in red , a *negative* line ($\Upsilon = -1$) is in blue whereas a *both* line ($\Upsilon \sim 0$) is in green . The color area in the panels represents the zone where a significant point must be situated to be attributed to the line (width of 0.1 Å) as well as the zone where the pseudo - EW was computed.

Our second class is based on the pseudo - EW. For each line with significant variation centered on wavelength λ_s as defined by the HARPS K5 catalogue, the area is computed in the spectrum ratio on the left ($\lambda_s - 0.05$Å) and the right of the line center ($\lambda_s + 0.05$Å). A width of 0.1Å appeared as a good choice for most of the detection peaks. The zone used for the pseudo - EW computation can be seen in the case of the 4375 Å FeI line in Fig.12 or the 4616 Å CrI line in Fig.13. The left area (A_{left}) and the right area (A_{right}), when summed, gives the pseudo - EW. We computed the asymmetric coefficient A_{left}/A_{right} for each variation line during the 47 nights and looked at the correlation of this variable with activity. Note that we removed outliers in the A_{left}/A_{right} variable by performing a 2σ clipping. The distribution of regression slopes is then divided in three tiertiles : the first one with negative slope represents the *negative asymmetry* whereas the last tiertile with positive slope forms the *positive asymmetry* group. A symmetric variation should produce

a slope null. Since, a symmetric flux variation is not displacing the barycenter of the line used to determine the RV, we will neglect this group. One advantage with this criterion is that even if the line center is not precisely determined, a symmetric variation still produces null regression, as it can be seen in Fig.16.

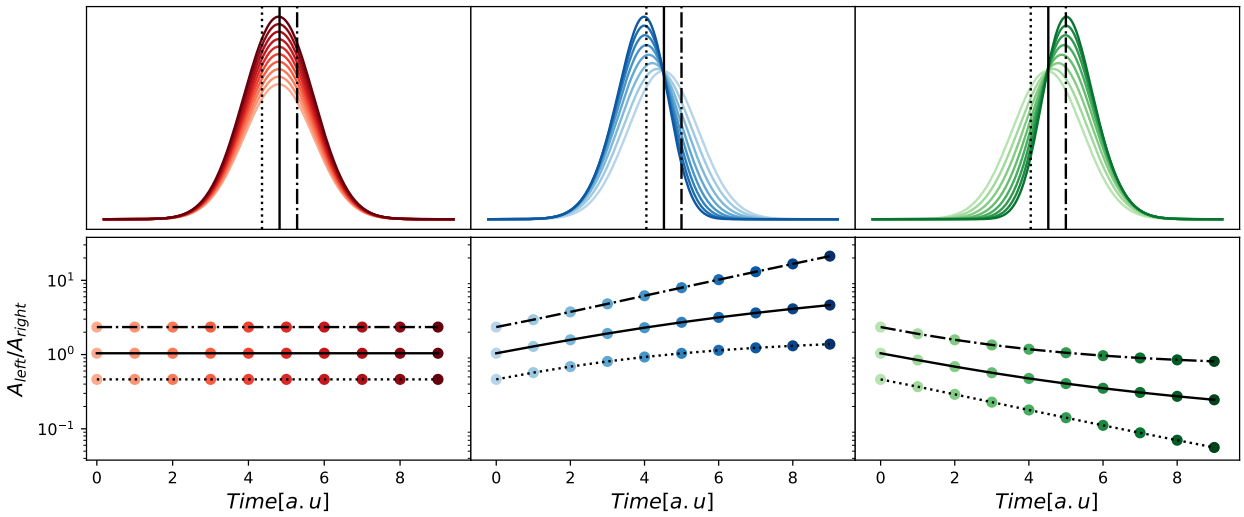

Fig. 16: Evolution of the A_{left}/A_{right} value as a function of time (or activity level) for three type of variations in the ratio profile. For each case, three different «center» are displayed : a line center well identified (*solid lines*), a line center miss identified on the left (*dotted line*), a line center miss identified on the right (*dashed line*). Symmetric variations can be identified by a regression slope null regardless if the line center is well identified. **Top :** Different modifications of a positive peak as the 4375 Å line (Fig.12) evolving with time or activity level. **Bottom :** Correlation graphics. **Left :** A symmetric variation is not changing the left over right area ratio. **Middle :** Same as left with the development of a left asymmetry (blueshift) which produces a positive regression slope. **Right :** Same as left with the development of a right asymmetry (redshift) which produces a negative regression slope.

We must precise that the asymmetric area parameter and the asymmetric height parameter (Υ) are two distinct classes that provide different informations and that cannot be compared individually when conclusions about RV will be presented.

4.3. « Zooprofile »

We simulated relevant lines variations to understand how the variations looks like in the "spectrum ratio space" and what values would take our two asymmetric parameters. The expected "zooprofile" is displayed in Fig.17. We assumed a perfect Gaussian line (the conclusion would not be very different for Lorentzian or Voigt profile) and studied how modifications on this line are translated in ratio profile.

The most simple change would be a depth modification (1^{st} row Fig.17), as detected for the 4375 Å FeI line. This modification is known for temperature sensitive lines like neutral metals (Gray 2005). It produces a *symmetric* ratio peak either *positive* or *negative*. A Doppler shift (2^{nd} row Fig.17) would induce a strong *asymmetry* creating a "*thunder*" peak. A third category (3^{rd} row Fig.17) would be FWHM broadening. The typical ratio shape is a "W" or a "M". Conclusion are similar for the case of rotational broadening which preserves the EW of the line (4^{th} row Fig.17). We precise that our criterion Υ is not well suited to categorised these two latter classes specifically since at maximum only the two highest peaks are consider and not three. Combination of the previous transformations can provide a more complex ratio spectrum as displayed for a depth modification and Doppler shift (5^{th} row Fig.17)

The last important transformation of a line is the modification of its asymmetry. Often, a line is simplified as purely described by a Gaussian, Lorentzian or their convolution profile which are symmetric functions with respect to the line center. However the reality is different, the center of a line is the flux coming from the most external layers of the photosphere whereas the continuum comes from deeper layers. The most external layer can be seen as matter « at rest » because the velocity convection vanishes at the surface whereas deeper layers are moving toward us because of convective blueshift (see section.2). The continuum and the upper part of the line are thus slightly blueshifted introducing a so called "C" shape. In Fig.18, we introduced as an example a "C" shape asymmetry parametrised with a parabola preserving the line center. Lines in the quiet photosphere present naturally "C" shape asymmetries due to stellar convection whereas "inverse C" shape has been found for active regions in three FeI lines (Cavallini et al. 1985). The tool used to measure lines' asymmetry is the bisector which is an imaginary curve that passes through the middle of the line profile at each flux level (red dotted curve in Fig.18).

We introduced an intrinsic asymmetries (6^{th} row Fig.17) using a parabola as first approximation. The asymmetry preserves the center of the line which is expected since the line's center is coming from the most external layer where the movement of convection is negligible. Asymmetries introduce in the ratio profile : a strong area asymmetry which may look very similar to a Doppler shift.

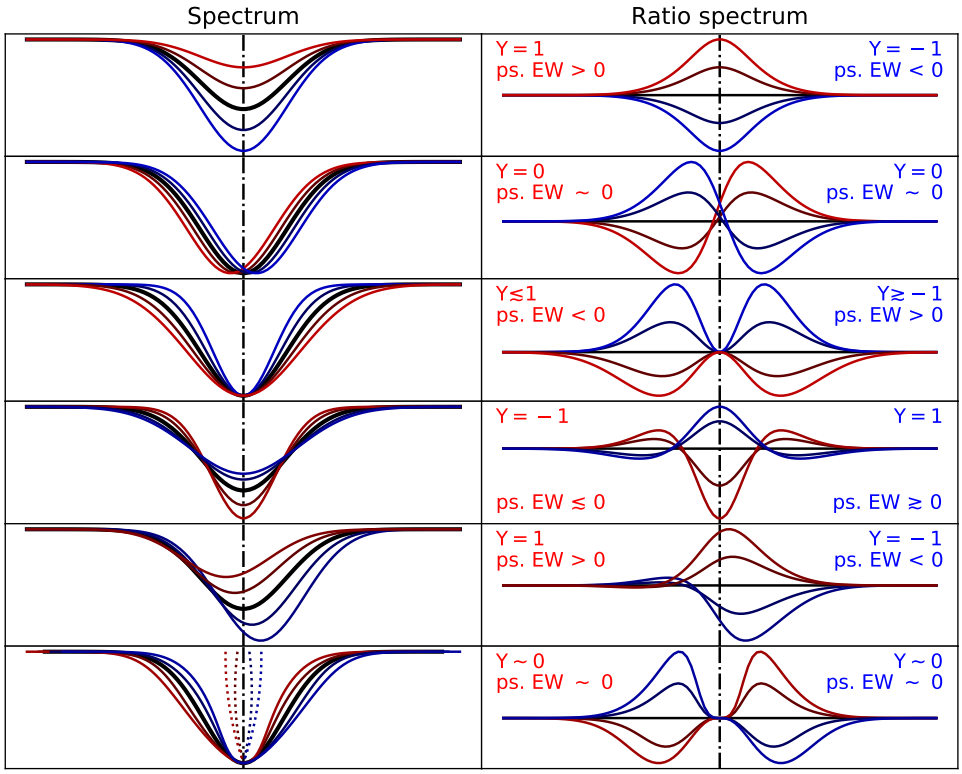

Fig. 17: **Left column** : Simulations of different transformations applied on a theoretical Gaussian absorption profile (*thick black line*). The initial line center is emphasised (*dashed dotted line*). **Right column** : Ratio spectrum of the transformed line profile normalised by the initial Gaussian profile. The Υ parameter and the pseudo - EW are indicated. **First row** : Depth modifications which should be flagged as *positive* ($\Upsilon = 1$) or *negative* ($\Upsilon = -1$). **Second row** : Doppler shift that should be flagged as *both* ($\Upsilon = 0$) lines. **Third row** : Broadening modifications in the wings. **Fourth row** : Rotational broadening modifications. It is actually not trivial to guess in which Υ classes such lines would fall, because it depends on the significance of the lowest peaks in amplitude. **Fifth row** : Combination of depth modification and Doppler shift. **Sixth row** : Different amplitude of line asymmetry as highlighted by the bisector (*dotted lines*). The center of the line is not affected. The ratio profile would be flagged as *both* lines.

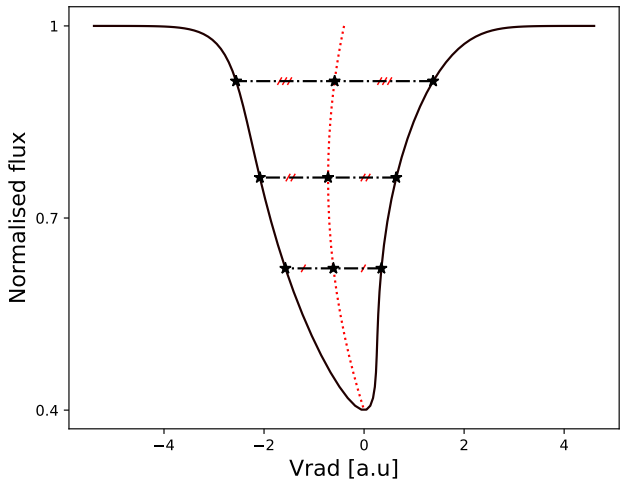

Fig. 18: Illustration of the geometrical constructions (***black*** *dashed-dotted lines*) to draw the bisector (*red dotted curve*) of the line (***black*** *solid line*). Here a "C" shape bisector reproducing the typical line shape formed in the quiet solar photosphere (Cavallini et al. 1985). The black stars markers are equidistant by construction.

4.4. Statistics on the selections

Once the groups are formed, we investigate if there are notable differences between them. For instance, are the individual radial velocities (RV_i) of a group better correlated with the radial velocity measured on all the lines (RV_m) than another one ? Our initial guess is that the *positive, negative* lines should have less effect on the radial velocities compared to the *both* ones. There are 5129 lines in the K5 catalogue, whereas we got the individual radial velocities for 5971 lines. This can be explained by the fact that the RV of individual lines are computed from the 2D HARPS spectra and the design of the spectrograph, with overlapping orders, implies that some spectral lines appear on two consecutive orders. We always keep the «twin line» with the smallest average error bar over the 47 nights. In the end, we got the RV_i for 4419 unique lines.

To check if a line is significantly affected by activity, we calculate the correlation between RV_i and RV_m. A strong correlation will indicate a strong dependence with activity as once the signal from α Cen A is removed, RV_m present strong variations due to activity (Dumusque 2014). We therefore compute the coefficient of correlation $\mathcal{R}_{pearson}$ between the RV_i and the mean radial velocity RV_m, the RV are binned by night of observation. We apply the same correction on RV_i and RV_m. First we remove the linear trend coming from the binary companion α Cen A after adjusting a sinus superposed with a linear drift (see Fig.19). The two stars exhibits a mutual revolution period of 80 years, because the maximum time between our observations was 80 days, the Keplerian signal can be approximated by a straight line and the residual contains mainly the activity signal. The use of a sinus is to account for the stellar activity effect and thus to prevent any bias when fitting the RV signal from binary. A 1.5 interquartile (IQ) clipping is then performed on the residues to remove outliers which could create false correlation. At the end only, the correlation was computed between the RV_i and the RV_m through a weighted linear regression (see the previous section 3.2 for an example Fig.8). RV_i were placed on the y axis because their error bars are always larger than the ones for RV_m.

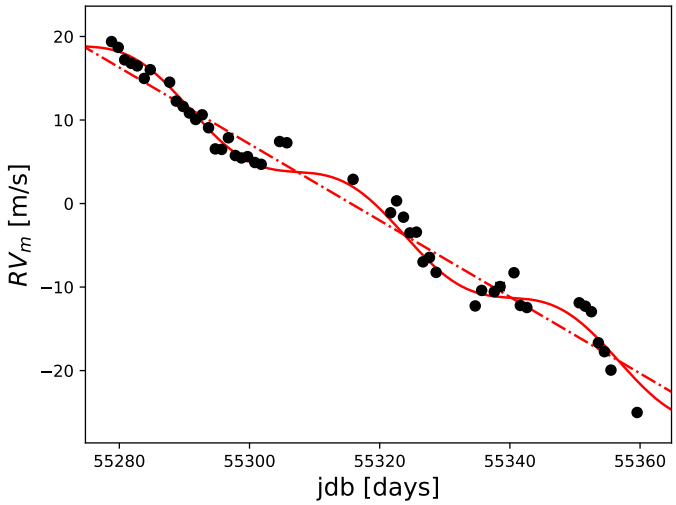

Fig. 19: 1-day binned average radial velocity RV_m of α Cen B (*black dots*) as a function of time (jdb - 2'400'000). A linear trend (*red dashed line*) induced by the companion star α Cen a has been removed after adjusting a sinus with a linear drift (*red solid line*). We note that the sinus is only a first approximation to the activity signal.

We displayed the Pearson coefficient comparing the group *no variation*, with the *variation* one (*positive*, *negative* or *both*) in Fig.20. Clearly, the normalised distribution of the *variation* lines show a strong excess of correlation compared to the *no variation* ones. There are still lines which seem to be categorised as not varying even though their RV are correlated with the activity, but at this point we must emphasise a bias in this ratio selection method more detailed in the next section. Outlines being that deep lines will exhibit most easily a strong peak that shallow lines. We can investigate if our initial guess is also satisfied inside the *variation* group itself. We must admit that we see no strong evidence of difference between lines expected to be more correlated with the activity (*both* lines) and those which are not (*positive*, *negative* lines). A slight excess is presumably present for the *both* and *asymmetric* ones for $\mathcal{R} \gtrsim 0.5$ but is still marginal. Strangely, for the asymmetric groups, since one group should be redshifted and the other blueshifted we could have expected a distribution centered on the negative $\mathcal{R}$ and another on the positive $\mathcal{R}$ but it is not the case.

A second coefficient of correlation is also computed between the pseudo - EW and the activity index $\log R'_{HK}$ (Fig.21). The three distribution of the *variation* lines are definitely not similar when compared two by two. The pseudo - EW of *positive* lines is strongly correlated with the activity as previously discussed for the case of the 4375 Å FeI line in Fig.12. In the opposite, a strong anticorrelation is observed for the *negative* lines. Since the majority of lines produced in the spectrum of α Cen B are absorption lines, these results show that *negative* lines are becoming deeper or broader when the star is active, whereas *positive* ones are becoming shallower (as highlighted in Fig.12 and Fig.13). Finally, a broad

distribution centered at 0 is observed for the *both* lines, coherent with the zooprofile (2^{nd} and 6^{th} row in Fig.17) where we see that the pseudo - EW is always more or less null regardless of the Doppler shift or the line asymmetry.

We describe the different composition of each classes in Table.1. We note first that the majority of *positive* lines are either *symmetric* or *positively asymmetric*, as well as the majority of *both* lines are *symmetric* or *negatively asymmetric*. Even if, the *negative* lines is the smallest sample, a net advantage is given for *negative asymmetry*. We remarked visually that all the lines in our spectrum were not necessarily reported in the K5 mask. In some cases, two peaks from different lines were gathered by mistake. And even the opposite critic could be reported, since we visually noted that some *both* lines can be mismatched because one of the peak were to far from the other or to far from the given line center (we remind that the line center is a possible source of systematics).

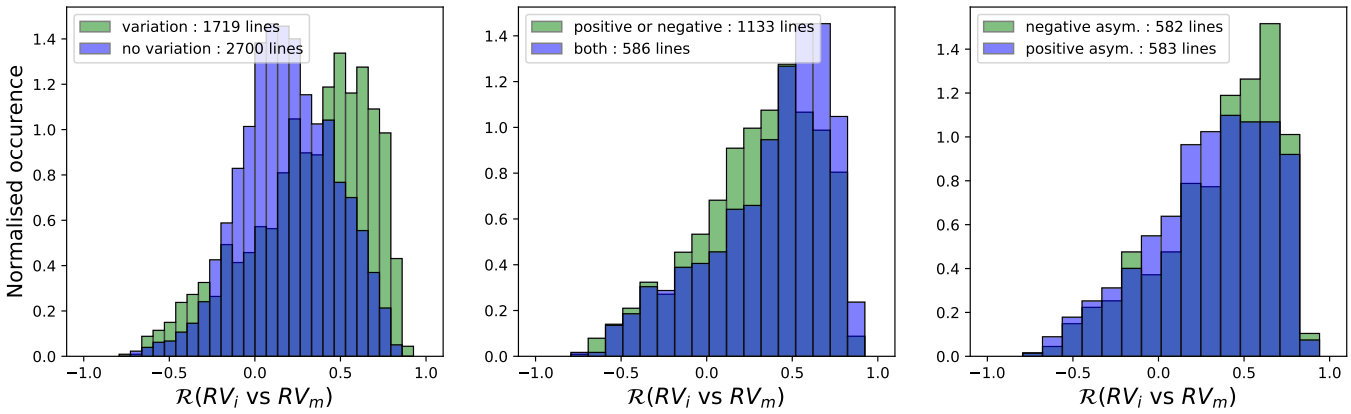

Fig. 20: **Left panel** : Histogram of the coefficients of correlation $\mathcal{R}$ between the RV_i and RV_m of the *variation* (*green*) and *no variation* (*blue*) groups. An excess is observed for $|\mathcal{R}| > 0.25$, illustrating that our method of selection based on ratio profile seemed to reach its purpose partially. There are still *no variation* lines showing strong correlation but it could be caused either by spurious correlations or by bias such that it is more difficult for shallowest lines to exhibit a peak (see section 4.5). **Middle panel** : Same as left for the *positive* and *negative* lines (*green*) versus the *both* (*blue*) lines. A slight excess presumably no significant is observed for $\mathcal{R} > 0.5$. **Right panel** : Same as left panel for the *negative asymmetry* (*green*) versus *positive asymmetry* (*blue*) lines. We see a slight excess, probably not significant.

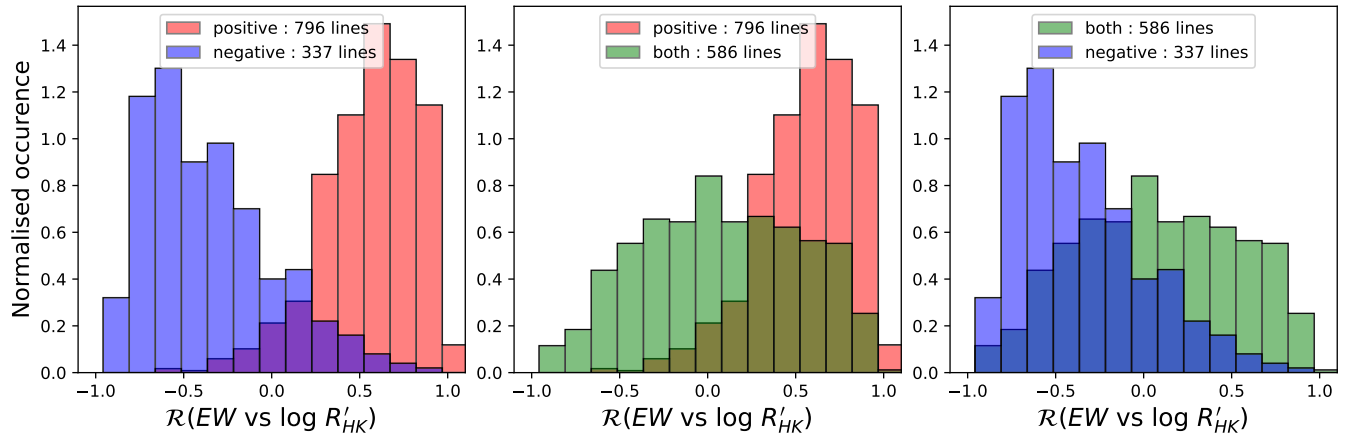

Fig. 21: Normalised distributions of the coefficient of correlation $\mathcal{R}$ between the pseudo - EW and log R'_{HK} of the different *variation* groups compared two by two. The three groups show clearly different distribution which is a good indicator that our groups are well formed. The *positive* (*red*) and *negative* (*blue*) lines show a strong opposite behaviour which can be explained considering that the pseudo - EW is signed. Therefore, a *negative* line, which becomes either deeper or broader with activity, increase the pseudo - EW towards the negatives. *Positive* lines are becoming shallower, increasing their pseudo - EW, such that positive correlations are observed.

Table 1: Composition of the classes and overlapping between Υ classes and the asymmetric groups. The pseudo - EW has not been computed for the no variation group.

	No variation	Positive	Negative	Both	Σ_{row}
Positive asymmetry :	-	378	107	98	583
Negative asymmetry :	-	114	162	306	582
(Symmetric)^a :	-	304	68	185	554
Σ_{col} :	2700	796	337	586	4419

Notes. ^{a} Symmetric lines are not studied later

4.5. Results on the radial velocity

We computed the mean radial velocity of the different groups (see section 3.2). The RV_i are grouped according to their "identity" (*variation, no variation, positive, etc.*), 1 day binned, and the average signal is measured with a weighted average :

$$< RV_i >= \frac{\sum_i RV_i \cdot w_i}{\sum_i w_i}$$

The linear trend from α Cen A is finally removed providing us different RV signals (Fig.22). In order to quantify if a signal is induced strongly or not by the activity, we started from the principle that, since the linear trend has been removed, the remaining fluctuations are mainly due to activity, thus if we achieve to mitigate the signal from the activity contribution, the rms should go down. On the opposite, if we are considering lines strongly affected by activity, the amplitude of the fluctuations should be larger and the rms larger in consequence. Let us precise that we used a weighted standard deviation to compute the rms. The RV averaged over all the lines, present a rms of 2.13 m/s.

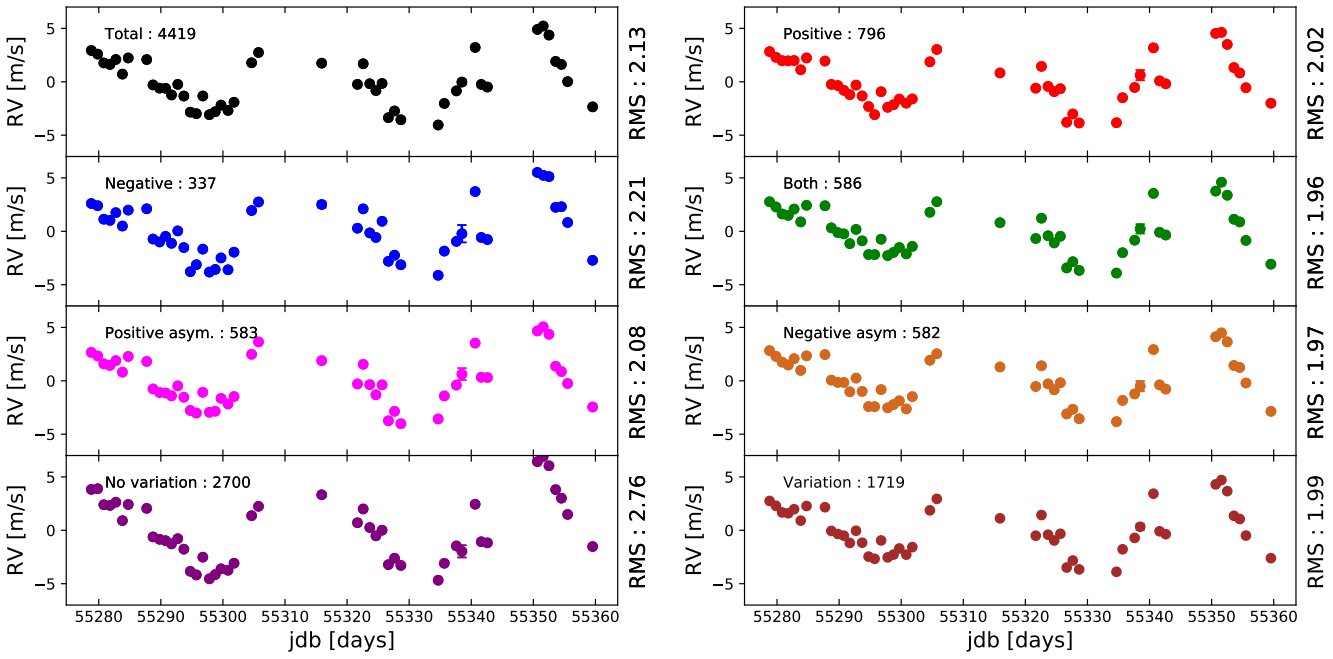

Fig. 22: The different RV signals of activity for α Cen B depending on our previous groups classification. The name of the different classes is indicated in the top left corner of each subplot (black : *all the lines*, red : *positive*, blue : *negative*, green : *both*, light purple : *positive asymmetry*, light brown : *negative asymmetric*, dark purple : *no variation* and dark brown : *variation*) as well as the number of lines in the different selections. Their « activity significance » is measured through their rms, shown on the right.

Clearly the results exhibit exactly the opposite behaviour than the one expected. *Both* lines show a weaker activity contribution than *positive* and *negative* ones, whereas *positive asymmetric* and *negative asymmetry* lines produce similar RV signals. However, the most concerning result is that the *no variation* group seems more driven by activity than the *variation* one !

To explain it, we point out that the method of studying spectra ratio is biased by the depth of the line. We drew it more explicitly in Fig.23 by representing two Gaussian in absorption : $\mathcal{N}(0,1)$, continuum normalised. The first Gaussian has a depth of 40%, when the second one is twice deeper (80%). We shifted both line by a distance equals to 0.33, 33% of the sigma width. The consequence is the following : in ratio flux, for the exact same shift, the deepest line exhibits a peak amplitude of 40%, compared to 10% for the shallowest one. Moreover, we noted that the positive peak observed from the thunder shape is necessarily higher than the negative one caused by the same bias (solid vs dashed curves). Hence, positive peaks emerge more easily that the negatives ones, coherent with the statistics number of lines positive and negative reported in Table.1. We end, by representing how the height of the peak (called its amplitude) is linked with the depth of the line for different Doppler shift. From this figure, we can conclude that shallowest line, independently from the shift or the distortion of lines' asymmetry, will never exhibits strong peak and thus will never be selected in our classes assumed to be linked with activity.

This result is also supported by the absence of shallow lines in our selections on the Fig.14 and the Fig.15, but above all by the histogram of the lines depth on the parameters plot Fig.24. By the way, this observation could kill two birds with one stone, explaining at the same time the presence of correlated line in the *no variation* class, which could be shallow lines, as well as the presence of uncorrelated line in the *variation* class. Indeed, for a very deep line the slightest flux variation will produce a peak.

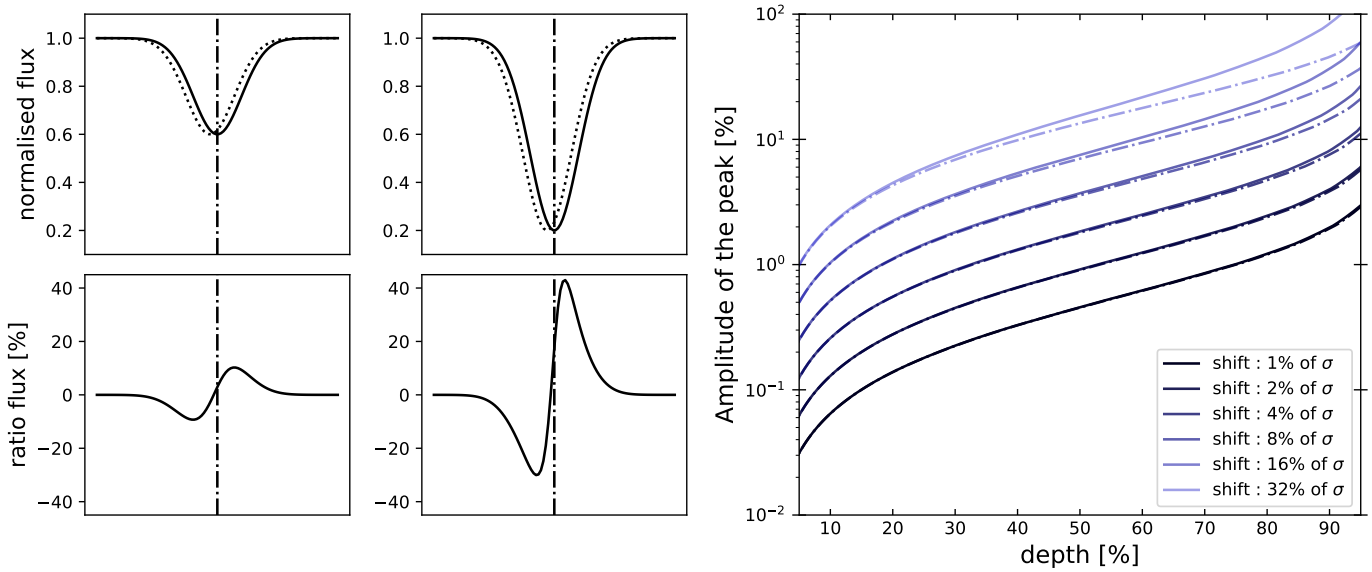

Fig. 23: Illustration of the bias of this method of selection through ratio profile. **Left part** : Two Gaussian absorption lines (depth of 40 and 80 %) shifted by the same distance (33% of the σ width). The shallowest line produces a ratio peak of 10%, whereas the 2 times deeper line, produces a 4 times higher peak amplitude. As a comment, the same bias is actually inducing a higher positive than negative peak in the thunder shape. **Right part** : Graphics of the positive peak amplitude (*solid curves*) and negative peak amplitude (*dashed-dotted curves*), in absolute since the y-axis scale is logarithmic, versus the line depth for different Doppler shift simulations. Shallowest lines have no chance to produce significant peaks and thus only deep lines will be selected through this method.

Also, from the parameter space distribution (Fig.24), it turns out that the Wavelength-Depth phase space seems to very well split the two groups. Indeed, we observe that a shallow blue line will be marked as a *no variation* line when a deep red line seems to produce more easily a peak in spectra ratio. We expect from activity some colour effect, meaning that some spectral range are more affected by activity (confirmed in section 5.4). But here we are wondering if again a bias is not at play. Assuming the same shift v for all wavelength, from the formula of Doppler shift equation (1), we can easily show that the RV shift of a 6000 Å line will be twice the one of a 3000 Å line :

$$\frac{v}{c} = \frac{\Delta\lambda}{\lambda_0} = \frac{\Delta\lambda_{3000}}{\lambda_{3000}} = \frac{\Delta\lambda_{6000}}{\lambda_{6000}} \Longrightarrow \Delta\lambda_{6000} = 2 \cdot (\Delta\lambda_{3000})$$

We remind that it is precisely for this reason that the lag of the CCF is not in linear but logarithm scale. The presence of bias by respect to the line depth and the frequency is definitely concerning, in particular as those two parameters turn out to be the most relevant parameters to predict if a line is or is not driven by activity (see section 5.4).

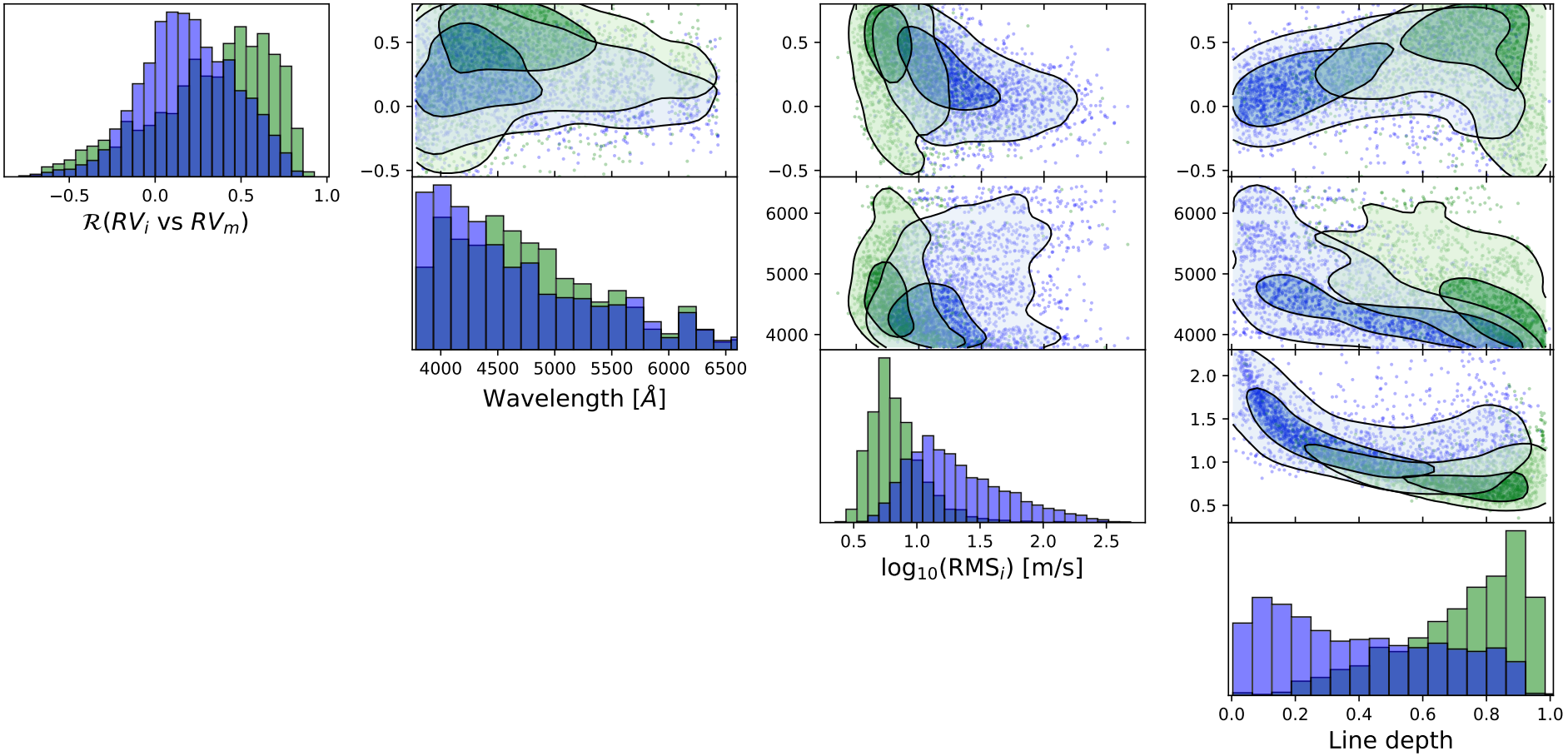

Fig. 24: Parameters phase space projections. (*Diagonal*) The normalised distributions of the *variation* (green) and *no variation* (blue) classes for 4 parameters : the coefficient of correlation $\mathcal{R}$ between the individual RV of the lines and the average signal RV_m, the wavelength, the RMS of the lines and the lines' depth. The bias in depth is clearly visible in the right-bottom corner. We also remark a clear separation in the Wavelength-depth scatter plot, showing that a shallow blue line is more likely a *no variation* line, when a deep red line has more chance to fall in the *variation* category. The unexpected behaviour observed in Fig.22 are linked with the observation that shallow lines (which show *no variation* in flux) exhibit a higher rms, this is even more true highlighting the fact that the graphic is not showing the rms of the line but its logarithmic value. The contour levels are at 1 and 2σ, but these sigma-values are for a 2D probability distribution, hence at 39% and 86% levels. For more informations about it, read : http://corner.readthedocs.io/en/latest/pages/sigmas.html

4.6. Attempts of corrections

Since the depth of a line is known, we could wonder if there is a possibility for us to correct this bias using the curves simulated in Fig.23. The first incertitude comes from the fact that our simulations are only considering a Doppler shift and thus should only correct the positive peak of *both* lines, even thought the depth bias is without doubt present for all the categories.

We normalised the ratio profile by a correction which is a function of the line depth. Indeed, from our simulations, we note that the positive amplitude A of thunder lines is a linear function of the shift $\Delta\lambda$ (proved by the fact that the curves on the right part of Fig.23 are equidistant) :

$$A(\Delta\lambda, d) = f(d) \cdot \Delta\lambda$$

The function f linked with the line depth (d) can be approximated by a polynomial function (we used a 15^{th} degree polynomial function). We remind that the height of the peak is entering at the beginning of our procedure in the significance criterion but never afterwards. Thus, the correction cannot be applied a posteriori on the selected lines but must be performed before the significance peak criterion, directly on the ratio profile.

We used our polynomial correction curve (based on the 1% of σ shift) to normalise the ratio profile but because this correction has been established for the positive peak of *both* lines, there are some speculations to use it. In particular, if the Thompson's method was biased towars deep lines, our method seems to reverse the issue such that closer to the continuum we are, more numerous noisy peaks emerge randomly. But this observation is maybe just supporting the idea that the precision in flux is not sufficient for such computations.

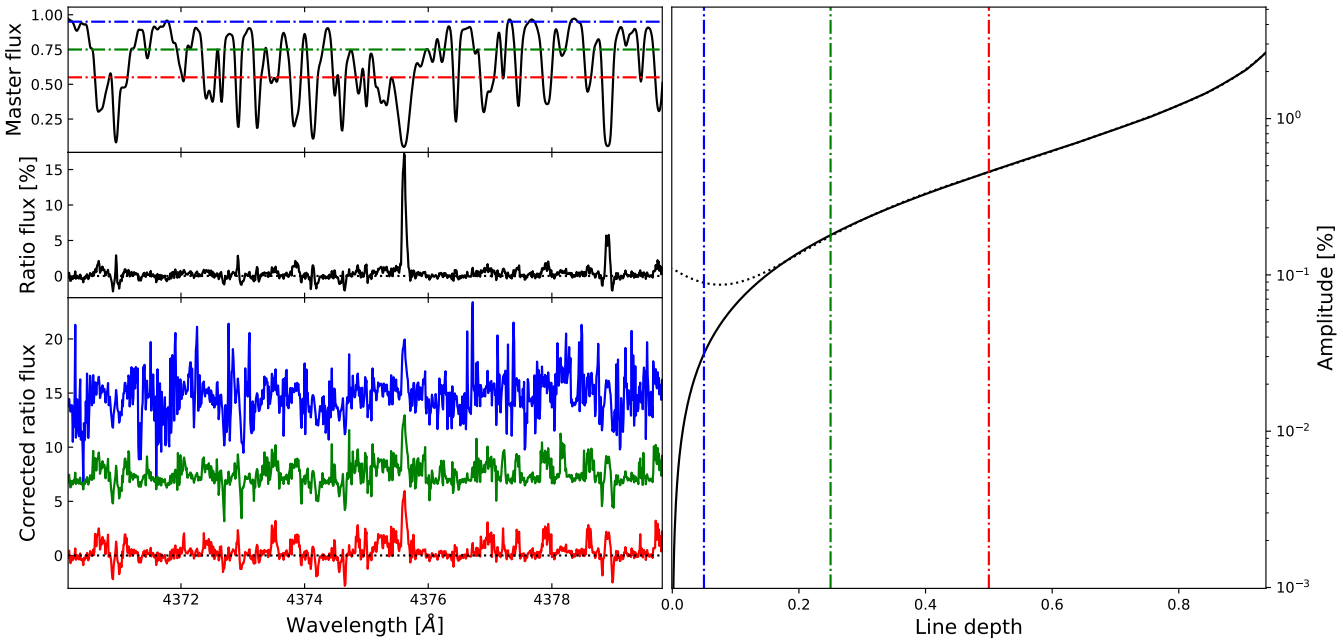

Fig. 25: **Right** : 1% of σ width Doppler shift curve from the previous graphic Fig.23. A 15 degree polynomial function has been fitted to established the correction (*black dotted curve*). The range for the fitting was restricted in depth between 0.15 and 0.95 to avoid the non realistic divergence around 0. **First raw** : Initial master quiet flux normalised. **Second raw** : Flux ratio (in percent) according to Thompson's method. **Third raw** : Flux ratio corrected by our polynomial correction with different offsets. Three distinct cutoffs in flux above which the correction has not be applied are displayed (red = 0.55, green = 0.75, blue = 0.95), the equivalence in term of line depth is also displayed on the right (*dashed-dotted lines*). The peaks reported by Thompson almost completely disappear like the 4379 Å one. However, our correction is highly dependent on the threshold chosen. Moreover, our correction is actually providing ratio peak where there is no line, because the noise is amplified.

We recomputed the significant peak selections using our correction without cutoff in flux (providing roughly the blue signal in Fig.25). We noted that several significant peak were actually lines not listed in the K5 mask of HARPS. Let us precise that because the number of lines was really insufficient with the 3σ criterion, we chose a less restrictive condition switching to a 2σ envelope. We counted 643 *positive* lines, 214 *negative* and 45 *thunder* shape (thus 902 *variation* lines). We did not compute the asymmetric criterion since our correction is disturbing the ratio profile shape. The small statistic size sample (except for *positive* lines) unable us to make an inter-group analysis, therefore, we chose to focus only on the

variation versus *no variation* lines.In Fig.26 we displayed the RV signal obtained, a slightly higher rms for the variation group is found, however it is difficult to be sure that 7 cm/s can be considered as significant and our assessment is that the precision in flux is not sufficient. We remind indeed that a ratio spectrum is inevitably noisier than the initial spectra. In consequence, our result is perhaps just supporting the idea that the information is wiped out, even thought we concede that a better correction than ours should be developed to be sure of it.

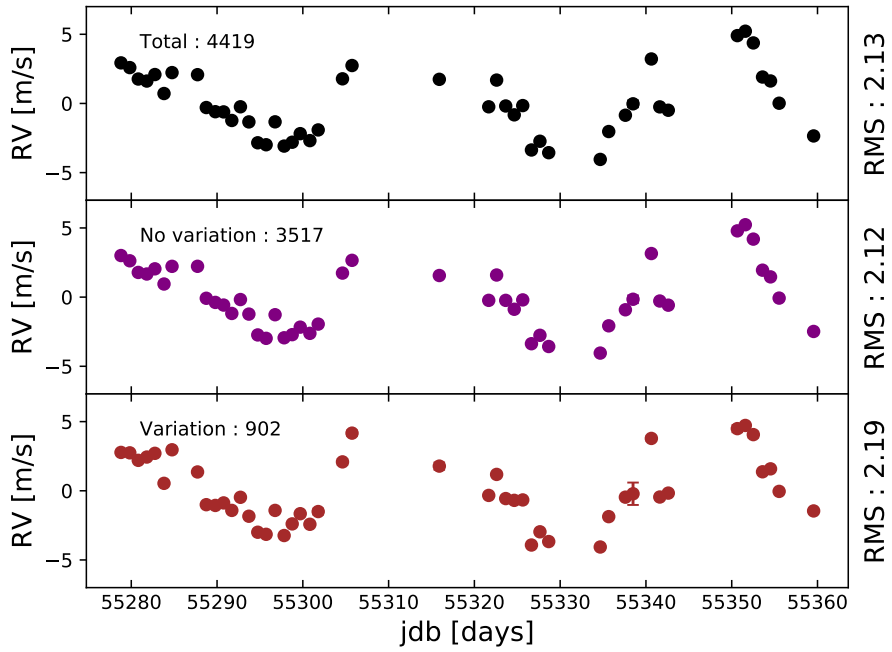

Fig. 26: Radial velocity signals for the *variation* and *no variation* groups as made for the Fig.22 but using our corrected ratio profile instead of the Thompson one. The variation group is now exhibiting a higher rms even though a such small difference is difficult to be considered as significant. The lack of difference observed is for us a good proof that the ratio profile washout the spectral information.

4.7. Conclusion

In conclusion, this method of selection has clearly missed its purpose. The initial idea was interesting, because with spectra ratio, each line is normalised by its own shape « at rest » and should in consequence highlight flux variations between high and low activity period. Moreover, some lines showed a clear modulation of their flux linked with the activity like in the case of the 4375 Å line (Fig.12). However, we noticed several difficulties and issues which make it difficult to use ratio spectra as a tool to probe lines inducing a strong RV effect. First, because each line is different (for instance in width), there is a priori no way to find a universal criterion available for all lines as we've done when we ask that a peak must be at a distance smaller than 0.05 Å from the center of the line. Alternatively, we can also wonder if the center is correctly identified and at which precision. Finally, our study shows that analysing spectrum ratio is more complex than looking directly at the spectra themselves as some information is lost in the division of one spectrum from another.

More concerning, we reported some dangerous bias of selection like the depth bias in Fig.23 (deep lines produce more easily peaks) and colour bias (red lines produces more easily peaks). We found no way to easily correct the former without producing other troubles. This aspect is even more compromising knowing that these two parameters are the most relevant to distinguish lines affected or not by activity as explained in the following chapter. Ultimately, from signal amplitudes shown in Fig.22 and Fig.26, we can conclude that this method is not suited to detect the lines the most sensitive to activity in terms of RV variation.

5. Classification based on the lines' physical properties

5.1. Overview

Until now, we tried to classify lines according to their observational morphology variation. But in fact, we already know which line is correlated or not with the activity since we know the individual RV_i and computed thanks to them the $\mathcal{R}$ coefficient of correlation with the RV_m. Of course, the coefficient of correlation is dependent on the precision we have for RV_i and for some lines the reliability is doubtful. But maybe that correlated lines share some common "physical footprint". Are they mostly produced by a peculiar element ? Are the atomic transitions sharing a similar range of excitation potential level ? The main idea is the following : two groups of lines are formed based on their correlation with the activity. For each line, several physical properties of the transition are known. An algorithm try to find out if some parameters are the "keys" to understand why a line is or is not correlated with the activity. Such objectives can be reached using machine learning algorithms.

5.2. Formation of activity sensitive and non-sensitive groups

We would like to form 2 groups : The first will be called the *correlated* and gather lines showing a signal similar to the activity. The other one will be called the *uncorrelated* and will be composed of lines less or not correlated with activity. Let us precise that here, to check the sensivity of a line to activity, we analyse the RV_i-RV_m correlation. This is because we know that there is not a one to one correlation between the activity signal in RV and $\log R'_{HK}$. However, this suppose that the signal observed in RV, in this case RV_m, is mainly affected by activity. For α Cen B in 2010, we are sure that the RV observed are induced by activity since a similar modulation is visible in the $\log R'_{HK}$ index.

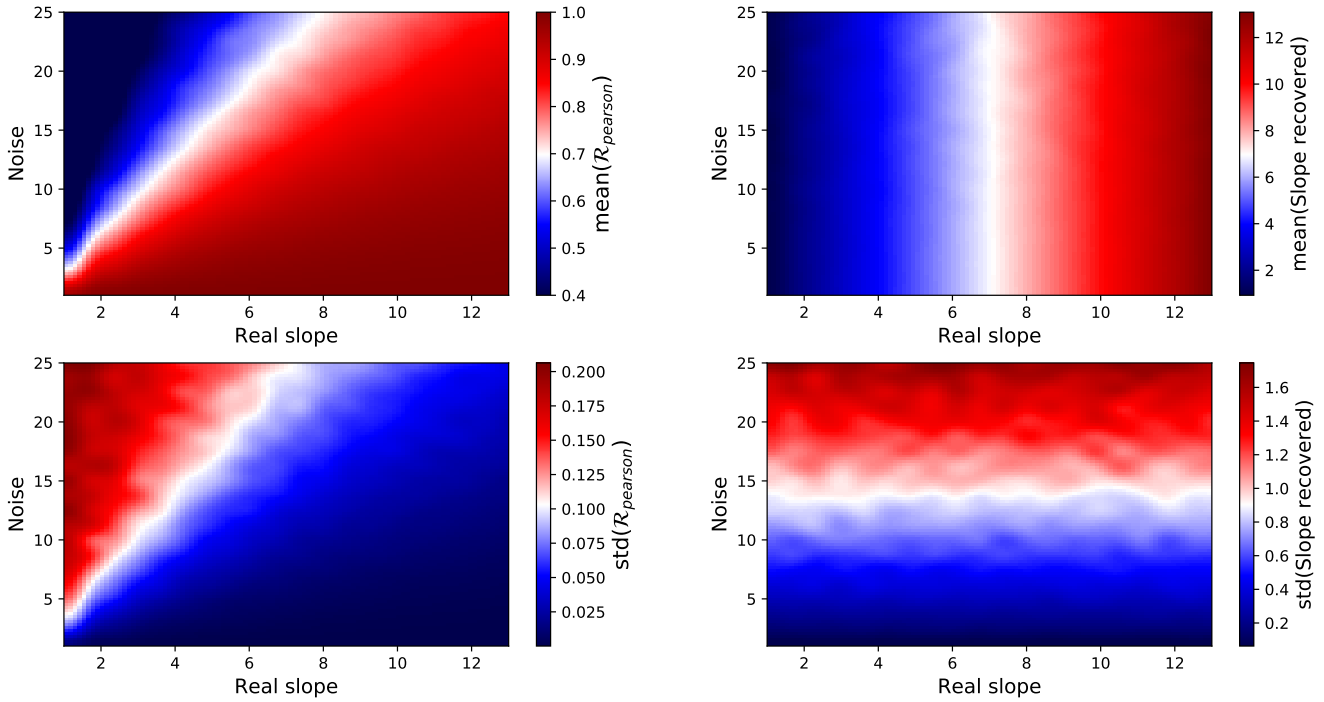

Fig. 27: Impact of the noise on the Pearson coefficient and the slope when fitting a straight line to data. For each combination of initial slope S_{real} and Gaussian noise level σ, we generate 500 different data sets and measured what is the recovered slope S_r and the Pearson correlation coefficient $\mathcal{R}$. This gives us for each couple of S and σ a distribution for S_r and $\mathcal{R}$. We then plot the mean and the standard deviation of those distributions on the top and bottom panels, respectively. The plots on the left show the effect of noise on the Pearson correlation coefficient $\mathcal{R}$, and the plots on the right, the effect of noise on the recovered slope S_r. Let's take the example of the simulations with an initial slope of 5 and a noise of 15. The 500 simulations give $\mathcal{R} \simeq 0.7 \pm 0.1$ and $S_r \simeq 5 \pm 1$. For the same real slope with a gaussian jitter of 25, we found : $\mathcal{R} \simeq 0.55 \pm 0.15$ and $S_r \simeq 5 \pm 1.6$. In conclusion, the estimated slope is not biased due to noise, whereas for the Pearson coefficient, we get smaller values than expected when the noise is increased.

The formation of the *correlated* and *uncorrelated* group is the key part of the analysis. This selection has to be made directly on the RV_i features and regardless of the physical properties of the line, otherwise it would biased afterwards the machine learning process. Fortunately, we know the RV_i of each individual line and thus we are in possession of a precious and purely observational information (see section 3).

We could think a priori that the coefficient of correlation $\mathcal{R}$ is well suited to indicate if a line is strongly similar to the activity signal, but after some investigations it turned out that the coefficient $\mathcal{R}$ is slightly biased depending on the noise level (jitter) of a line, whereas the slope of the regression S_r appeared like a more robust parameter. To prove it, we first generated a perfect line for x varying between 0 to 10 composed of 25 evenly spaced points (the slope S is a parameter). On this synthetic data, we added different levels of Gaussian noise $\mathcal{N}(0, \sigma)$. Because the jitter is by definition «noisy», there is a probability that the points deviate strongly from the initial line, that is why we generated 500 independent realisations with the same conditions (same slope S and level of noise σ). The slope S_r given by the correlation and the $\mathcal{R}$ are saved for each simulation. At the end, we recorded the mean and standard deviation of the slope and $\mathcal{R}$ distributions for a fixed combination of initial slope S and noise σ. This simulation has been made for different combinations of noise levels and slopes (Fig.27).

It turns out from this graphic that the pearson correlation is sensible to the jitter injected in the perfect line (for a fixed real slope, the pearson coefficient, in average, depends on the noise). On the opposite, we note that the slope S_r recovered by the regression is in average exactly the initial slope, regardless the degree of noise injected. The jitter has for effect to naturally spread out the distribution of the slope recovered, but without biasing its mean. We can easily understand why the slope is unaffected by Gaussian noise when the Pearson coefficient is. Gaussian noise will push symmetrically points above and below the initial line such that in average the slope is not disturbed. In fact, we do not even need the symmetric assumption to get the same result. An asymmetric distribution of noise would only impact the intercept of the line not its slope. The only assumption needed is that the noise is independent of the point. On the contrary, noise breaks the « line shape », measured by the $\mathcal{R}$. For this reason, our main parameter to form the group is the slope of the correlation and not the Pearson coefficient[1].

Then, we can wonder if a line exhibiting a strong temporal variation with high error bars should be considered, our fear being to contaminate our selection and to confuse the machine learning algorithm. For this reason, we will also use the rms of the line (RMS_i) defined as the standard deviation of the detrended RV_i and the median of the error bars (photon noise) in our criteria. The ratio between the RMS_i and the median of the error bars provides a kind of signal to noise ratio (SNR_i). For the *uncorrelated* group, a maximum RMS_i is fixed because a line without correlation is not necessarily exhibiting a flat RV signal (for a sinus, a signal shifted of $\pi/4$, has no linear correlation with the initial signal but present a circular correlation). The criterion to form the groups are explicitely written in Table.2.

Table 2: Criterion to form the groups *correlated* and *uncorrelated* based on 5 parameters. For the *correlated* : A lower boundary in slope (S^{min}_{corr}) and in SNR (SNR_{min}). For the *uncorrelated* a lower and upper boundaries in slope (S^{max}_{uncorr} and S^{min}_{uncorr}) as well as an upper boundary in RMS (RMS_{max}).

Category	Criterion
Correlated	$S_i > S^{min}_{corr}$ & $SNR_i > SNR_{min}$
Uncorrelated	$S^{max}_{uncorr} > S_i > S^{min}_{uncorr}$ & $RMS_i < RMS_{max}$

In Fig.28, we display an example of such selection, the criteria were : $S^{min}_{corr} = 2.1$, $SNR_{min} = 2$, $S^{min}_{uncorr} = -0.55$, $S^{max}_{uncorr} = 0.55$ and $RMS_{max} = 20$ m/s. This selection was chosen for the test session of the machine learning (see next section) because the two selections seem correctly formed and with roughly the same number of lines. Again, just for this analysis, we suppressed the twin lines with the highest median error bars. It avoids, for example, for twins to fall in different category and thus to confuse the machine learning algorithm (because twins line by definition have the same physical properties except the pixel and order location on the detector). The resulting groups are displayed in Fig.28.

We note directly the powerful advantage to possess the individual RV_i allowing to mitigate activity signal, down to 89 cm/s using the *uncorrelated* lines. We precise that a linear trend have been removed to reach this value. It may sounds strange to find out a linear trend since we removed from the RV_i the effect of α Cen A, but it is possible that the long-term magnetic cycle influences differently different lines selections. We can be dubious of the utility to have a such restrictive number of lines in particular if we want to studied other stars where the signal to noise is lower and thus the use of the maximum of lines is recommended. This purpose can be reached using machine learning algorithms.

[1] Another presumption to avoid the Pearson coefficient is that the RV_m are by definition the weighted average RV signal. If a line contributes a lot in this signal, because of a great precision (high weight), the $\mathcal{R}$ will be high. So the risk for a criterion based on $\mathcal{R}$ is actually to select lines based on their weight.

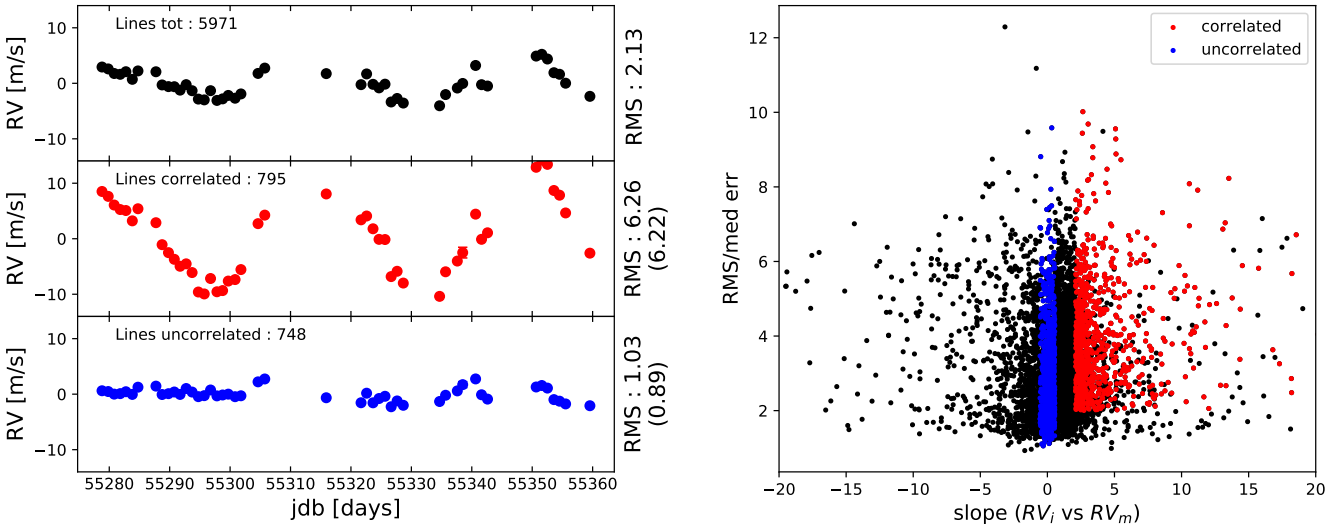

Fig. 28: **Left** : RV using all the lines (**black**), the *correlated* (red) with a regression slope $S_r > 2.1$ and a «SNR» higher than 2, the *uncorrelated* (blue) defined by $|S_r| < 0.55$ and a rms smaller than 20 m/s. An attentive reader would remark that the number of total line is more important than in Fig.22. The explanation come from the fact that for this subplot, we do not suppress the twins lines to obtain a RV that is similar to the one derived with the standard CCF (they were removed for the two below however). The rms is identical to Fig.22, since lines with high error bars mostly do not contribute in the signal because of their lower weights. The rms of the sub-selection (still a weighted standard deviation) are displayed on the right. In parenthesis, we display also the rms of the linear detrended signal. **Right** : Parameter space for the selections. The most restrictive condition is applied on the slope from the correlation with the RV_m. Others constrains can be applied on the rms, the photon noise of a line or even the ratio of both (kind of SNR).

5.3. Machine learning

Machine learning algorithms are powerful tools to take advantage of the progress and abilities of computer to «learn» and the predict outcomes in general cases (Lecun et al. 2015; Baldi et al. 2016) or in astrophysics (Huang et al. 2018; George & Huerta 2018). Let's introduce a metaphor to explain the basic idea hidden behind this term. Today we engaged a new postman for a very difficult task, his job is to guess if a letter is an express letter (called "A") or a classical letter (called "B") for the letters whose the stamps was unstuck accidentally (such problem is called a *binary classification*, but we could make the same exercise with more than two classes). Of course, our postman has neither superpower, nor divination skills and need to be trained to distinguish between letter "A" and letter "B". So the first step is the *training session*.

We brought two piles of letters : one on which the stamps « express » is still present and another with the « classical » stamps. We give to the postman, some rules or properties that he can use to distinguish the letters "A" from "B". For instance, the weight of the letter, its color, its shape, its size or even its state. Then the postman looks each letter and tries to see if some properties reflect the identity of the letter. Perhaps than all yellow letters are "A", so at least one aspect of the color properties can help to distinguish the two classes. Maybe than most of letters heavier than 100g and squared are « classical ». The postman do not tell us precisely what kind of computations he is performing, however he gives us the importance of each property helping his decision. When the postman has finished to learn, we must test it to be sure that his job will be correctly done. So the second stage is the *test session*.

In order to test him, a new bag of N letters is given to him. This time the stamps were removed but WE know if a letter is "A" or "B". We ask the postman to make predictions on each letter. Once he finished, we can compute his score through two fundamental parameters : the *recall* and the *precision*. There is initially N_A letters in the "A" - class and N_B letters in the "B" - class from the bag. The postman predicts P_A and P_B, respectively "A" and "B" letters. Some are good predictions, whereas other are false. There are $C_{A,A}$ "A" letters correctly predicted and $C_{B,B}$ "B" letters correctly identified. We can display all these informations in a compact form through what is called the *confusion matrix* (Table.3) which is a matrix illustrating and comparing the initial classes and the one predicted by the postman.

If this matrix is diagonal, the postman performed the best job possible. If the matrix is anti-diagonal, we should definitely fired him or at least let him train with a larger sample or others parameters, but in fact this scenario is certainly unlikely, because if the postman has failed the training sessions, he will attribute randomly the letters in the two categories, and thus we should find an homogeneous matrix (see annexe section C.1). The recall R of the class j is defined as :

$$R_j = \frac{C_{j,j}}{\sum_i C_{j,i}} = \frac{C_{j,j}}{N_j}$$

Whereas the precision Π is summed along the column :

$$\Pi_j = \frac{C_{j,j}}{\sum_i C_{i,j}} = \frac{C_{j,j}}{P_j}$$

It is important that both the recall and the precision are close to 1. Indeed, let us imagine that our postman decide to predict all the letters from the bag as "A" letters (see section C.2). It will provide a very impressive recall of 1 but a very bad precision (both recall and precision of the "B" class will be bad by the way). A useful value to combine the recall and the precision is the F^1 score which is the harmonic mean of the two previous quantity :

$$F_j^1 = \frac{2}{\frac{1}{R_j} + \frac{1}{\Pi_j}} = \frac{2\Pi_j R_j}{R_j + \Pi_j}$$

Again, all previous informations on the test session can be found in Table.3. Let us finish this short presentation by introducing others useful parameters as the *accuracy* A or the *error rate* E of the test, based on the trace of the matrix :

$$A = \frac{C_{A,A} + C_{B,B}}{N} \qquad E = \frac{C_{A,B} + C_{B,A}}{N} = 1 - A$$

Table 3: Confusion matrix summarising the results, extracted from the test session of the machine learning, completed by the different scores. The recall R and the precision Π allow to quantify the achievement of the machine learning process. The F^1 score (bottom right in the table) represents the mean harmonic of both the recall and the precision and can be computed for the "A" and "B" groups.

True \ Predicted	A	B	Σ_{row}	R
A	$C_{A,A}$	$C_{A,B}$	N_A	$C_{A,A}/N_A$
B	$C_{B,A}$	$C_{B,B}$	N_B	$C_{B,B}/N_B$
Σ_{col}	P_A	P_B	N	
Π	$C_{A,A}/P_A$	$C_{B,B}/P_B$		$\frac{2\Pi_j R_j}{R_j+\Pi_j}$

In the annexes (section C.1 and section C.2), we present two relevant cases of degenerate and thus bad algorithms which are the *random choice* and *overpredictions*.

5.4. Relevant parameters according to machine learning

The next question to ask is : Is there any specific physical properties that distinguish the *correlated* and *uncorrelated* lines ? We saw that the two selections are similar in number (roughly 750 lines), but when the first rms reaches 622 cm/s the second drops at 89 cm/s, barely seven times smaller. Could we found some physical parameters which explain our selections ? We will use two machine learning algorithms for this purpose : the *logistic regression* and *XGBoost*. The logistic regression is the most simple linear model used to predicted binary classification. It consists to place a $N-1$ dimensional plan in a N dimensional parameter space to separate the two groups. The XGBoost (let's say for simplicity XGB) predictions are based on decision trees that the algorithm develops (Fig.29). The maximal depth of the trees allowed was 3 levels, whereas 100 parallels trees were established[2]. As we will see, the logistic and XGB algorithms are working differently, providing different results and scores of predictions.

First, we must say to our algorithm which parameters are important to consider. We choose 16 parameters, 13 based on atomic physics given by the VALD catalog which records many atomic and molecular transitions (Heiter et al. 2008). These parameters are general but some are either entering into account for the line stellar depth formation (parameters 6 to 10) or linked with a sensitivity to the magnetic fields (parameters 11 to 13). We shortly present them, the reader who may want more details can refer to Piskunov et al. (1995) :

1. The central wavelength of the transition
2. The line depth : Depends on the optical depth and on the concentration of atoms involved in the transition
3. The number of contaminating lines : A line is sometimes the superposition of many close transitions

[2] http://xgboost.readthedocs.io/en/latest/python/python_api.html

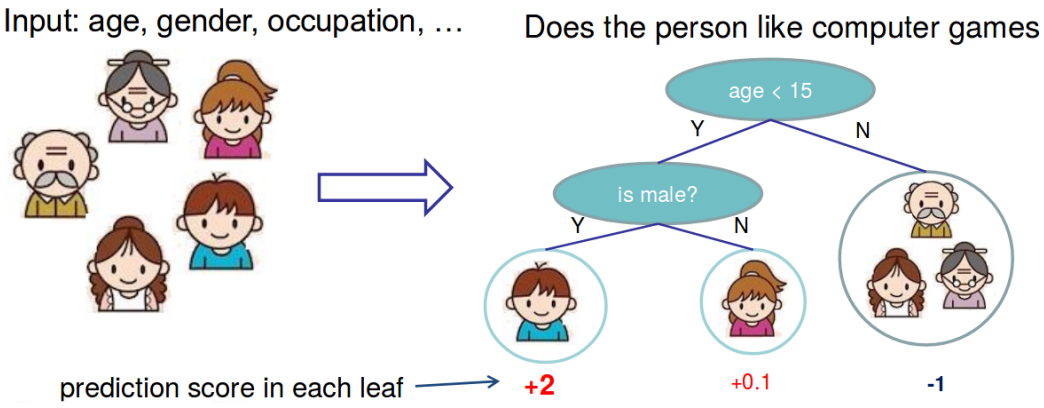

Fig. 29: Example of a decision tree classification to predict if an unknown likes to play on computer games. The algorithm makes its prediction on several parameters (here age and gender) because it learned that little boy were more likely gamers than old people. Let us precise that XGB is actually not only using one tree to make its predictions, but one hundred at the same time (*Crédits : http://xgboost.readthedocs.io/en/latest/model.html*)

4. The element
5. The ionisation level : the degree of ionisation of the element
6. The lower excitation level E_{low}
7. The logarithmic oscillator strength f times the weight g of the parent level $log(gf)$: enters into the central line opacity
8. The logarithmic radiative damping constant Γ_r
9. The logarithmic Stark damping constant Γ_s
10. The logarithmic Van der Waals damping constant Γ_w : the three damping constant provides the natural width of the Lorentzian profile).
11. The lower Lande factor
12. The mean Lande factor : Lande factors are needed to take into account the Zeeman splitting effect which is the lift of degeneracy of hyperfine atomic structure level in presence of magnetic fields.
13. The J_{low} quantum number of the total angular level of the lower energy level : The number of level after the splitting from Zeeman effect is a function of the angular quantum number J, higher is J, more hyperfine levels exist. Studies have observed molecular lines splitting in sunspots sensitive to the magnetic field intensity linked to this lift of hyperfine structure (Berdyugina et al. 2001).
14. The FHWM_RHK slope : the slope of the correlation between the $\log R'_{HK}$ index and the FWHM of the line.
15. The pixel position along the orders : pixels are numbered from 0 to 4096.
16. The coefficient of correlation $\mathcal{R}(RV_i$ vs $RV_m)$: we note that this factor is slightly biased with respect to presence of jitter on the RV_i.

We admit that the last three parameters are not physical but observational or instrumental. In any case, they are definitely of a great help for the algorithm as shown later. We checked that in the parameters selected, we did not have any strong correlations (full pair parameter plot in the last two pages of the annex). Any strong correlations between the parameters can perturb machine learning algorithms. Ultimately, we removed possible outliers from our two selections performing a 1.5 sigma (IQ) clipping on the 16 parameters. Because the standard deviation is more affected by outliers that the interquartile (IQ), we used this latter in the clipping. Only the 9^{th} until the 14^{th} parameters were concerned by the clipping, in extendo : the Lande factors, the FWHK_RHK slope, the Stark, Waals and radiation damping constants. At the end, we get 600 and 619 respectively *correlated* and *uncorrelated* lines in our selections on which the machine learning algorithm can work.

A stratified Kfolding composed of 3 groups is then performed to create the train and test samples. This means that we separated our initial sample of 1219 lines (600/619) into 3 different groups preserving the initial proportion of lines *correlated* and *uncorrelated* in each sub-group. Then the groups 1 and 2 are put together to form the training sample and testing occurs on the 3^{rd}. Once finished, the groups are permuted and the training is done on the 1^{st} and 3^{rd} put together whereas testing is performed on the 2^{nd} group, and so on for the last permutation. This method allows to be sure that the score of the test is not due to a specific combination of lines or low statistical uncertainties. In order to even more prevent such issues, we reproduced the procedure 200 times (each time with a new «seed» in the random generator). At the end, we are in possession of 600 simulations. For each simulations, we registered the noteworthy values like the parameters order of importance, the precision, the recall, for both the logistic regression and the XGB algorithm.

We display the parameter coefficients in Fig.30, higher are their values, more relevant are the parameters. To be able to compare the relative importance of the different coefficients in our logistic regression, we multiplied each coefficient

by the standard deviation of the parameters they depend on.. Moreover the coefficients are signed, the sign giving the direction of the gradient in the different plan of the parameter space. For instance, the fact than the $\mathcal{R}$ and line depth coefficients are opposite in sign show that while we must increase $\mathcal{R}$ to found lines *correlated*, the depth must decrease. Also the XGB coefficients are normalised such that the sum is 1.

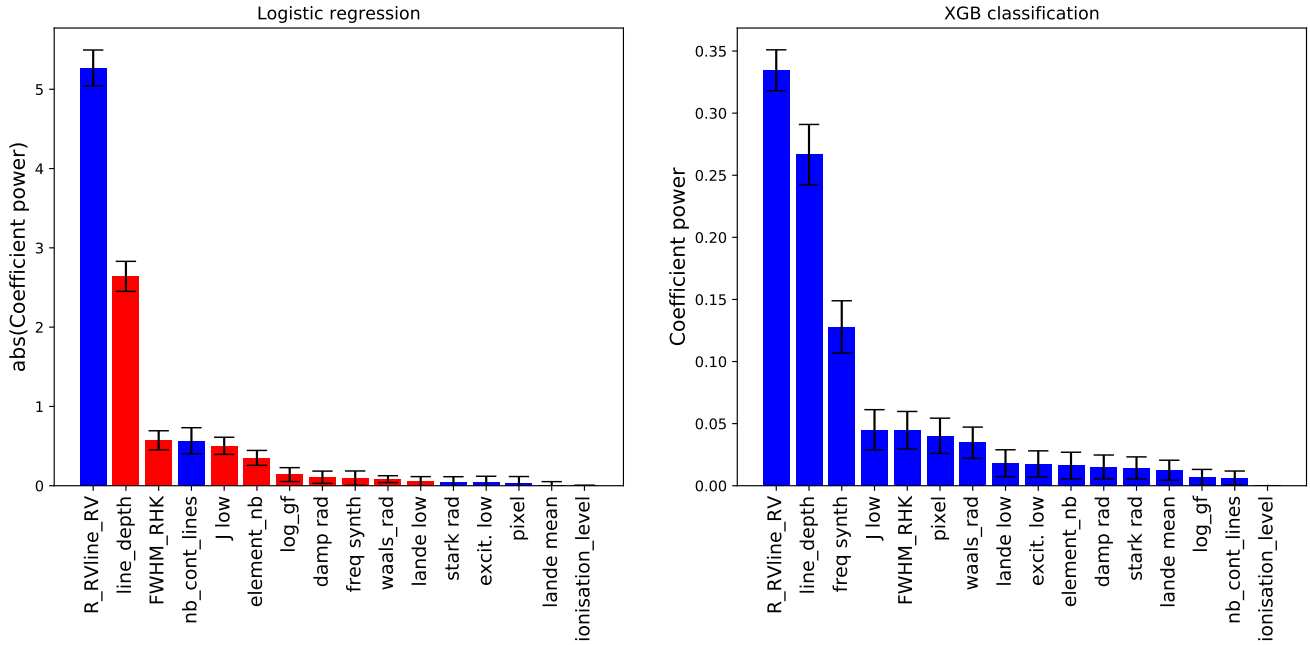

Fig. 30: Median of the parameters coefficient power for the 600 simulations. The error is given by the standard deviation of the obtained distribution. **Left** : Logistic regression results. Coefficients can be either positive (*blue*) or negative (*red*). The most dominant parameter dimensions are the $\mathcal{R}$ and the line depth. **Right** : XGB classification results. The podium is composed of the same parameters than for the logistic regression except that the wavelength seems this time more important than for the first algorithm. The order itself depends on the available parameters, we show in annex that after suppressing the observational parameters ($\mathcal{R}$ and the FWHM_RHK slope), the coefficients values and ranking are slightly reordered.

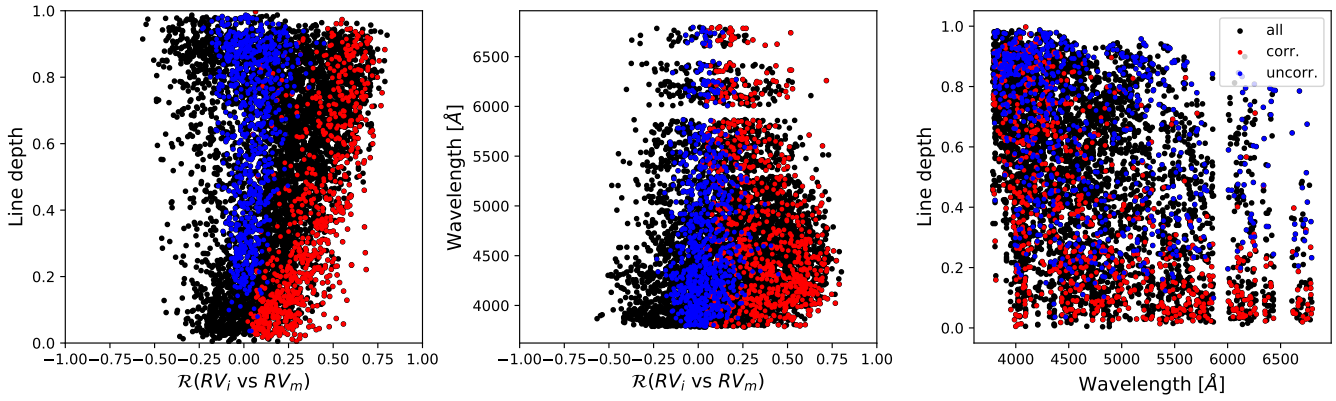

Fig. 31: Parameters plans of the three most important parameters according to XGBoost. The *correlated* lines are in red, the *uncorrelated* selection in blue , the others in black. **Left** : $\mathcal{R}$ versus depth plan. A very clear separation distinguish the two groups and explain why these two parameters are the dominant ones. A vertical separation give a greater importance to the parameter on the x-axis. **Middle** : $\mathcal{R}$ versus wavelength plan. **Right** : wavelength versus depth plan. XGB has understood something related with the wavelength, a color effect of the activity which is already known or expected (Huélamo et al. 2008). The absence of the wavelengths for the logistic algorithm is highlighting the fact that the two algorithms are working differently.

We explicitly display the 3 parameters plan of the most important parameters according to XGB (Fig.31). Clearly we distinguish a kind of barrier separating the two groups in the graphic depth vs $\mathcal{R}$. This barrier is mostly vertical, explicit computations giving for the slope of the frontier ~ 2.5 (meaning that the $\mathcal{R}$ is slighlty more important that

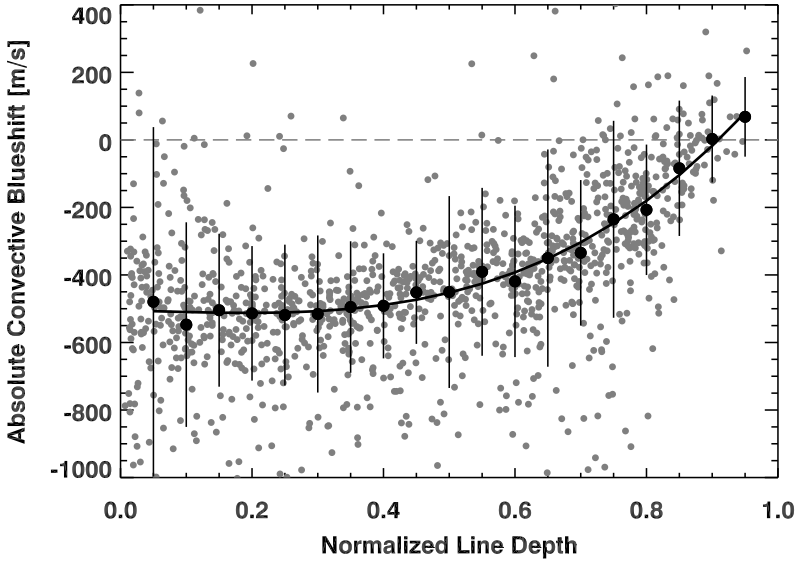

Fig. 32: Absolute convective blueshift for FeI lines measured for the Sun as a function of the line depth. (*Crédits :* Reiners et al. (2016))

the depth according to a logistic regression). The value of the slope can be recovered taking the ratio of the two power coefficients. Also we visually understand why the coefficient for the two parameters are opposite in sign. To reach one of the zones, we must either go on the top left or bottom right corner, meaning that to follow the gradient (if we imagine that blue points are a valley and red ones are up a hill), we must variate our parameters in opposite direction. Another interesting observation is to note that such significant barrier is only visible in the depth-$\mathcal{R}$ parameters plan. It is a particular configuration, since in our case only two parameters dominate the logistic algorithm. In a more general case, when many parameters are relevant at a same level, projections unable the visibility of such separation.

Line depth is an important feature to differentiate between lines strongly affected by activity and lines insensitive, since the magnetic fields inhibit the convection. As shown in Reiners et al. (2016), deep lines have an absolute convective blueshift of 0, while shallow lines $\sim$ 500 m/s. This is because shallow lines are formed deeper in the star, where convection velocity is higher. Thus shallow lines will be strongly affected, while deep lines not because they already have a null convective blueshift.

The comparison of the coefficients allow to point out some structural difference between the two algorithms since XGB clearly remark that the wavelength of the transition help its decision. XGB is known to « learn higher order of interactions between features »[3], as it can classify non-linear problems indicating perhaps that the color effect of activity is subtle but well present. Moreover, for XGB, the line depth and the $\mathcal{R}$ are roughly at the same degree of relevance.

We can also go back briefly on the Thompson results to better understand why this method did not work. We can find the same parameter plan in the top-right corner of the Fig.24 (axis were permuted). We remind that the ratio profile method was biased with the depth on the line (see histogram Fig.24), in other words, the method introduces in the first parameter plan an artificial horizontal frontier not physically involved to form the groups *variation* and *no variation*. But clearly, an horizontal frontier has no way to correctly separate our groups *correlated* and *uncorrelated*. In addition, since the *variation* group contains a majority of deep lines, shown to be less affected by activity, this explains the first surprising behaviour observed in Fig.22 of the higher rms for the *no variation* group compared to the *variation* one.

We have already a good idea of the most important parameters, unfortunately the $\mathcal{R}$ is not a physical property of the line and there is actually no guaranty that for all stars, the RV_i can be computed and sufficiently precise to allow an individual regression computation. As a consequence, we would like to determine the most important physical parameters. The problem is that, except the evident importance of the line depth, the others physical parameters happens bunched together. We thus perform a new ranking based on the occurrence frequency. For each simulation, we select the 8 most influent parameters over the 600 simulations (Fig.33).

A very good agreement is found between the two algorithms for the occurrence of the 4 best parameters. We remark however some strong disagreements for some parameters as the number of contaminating lines and the element type (= element number), which both seem to be used at some occasion by the logistic regression but never by the XGB classification. It highlights once more that both algorithms work according to different internal mechanisms. The consistence

[3] https://homes.cs.washington.edu/ tqchen/pdf/BoostedTree.pdf

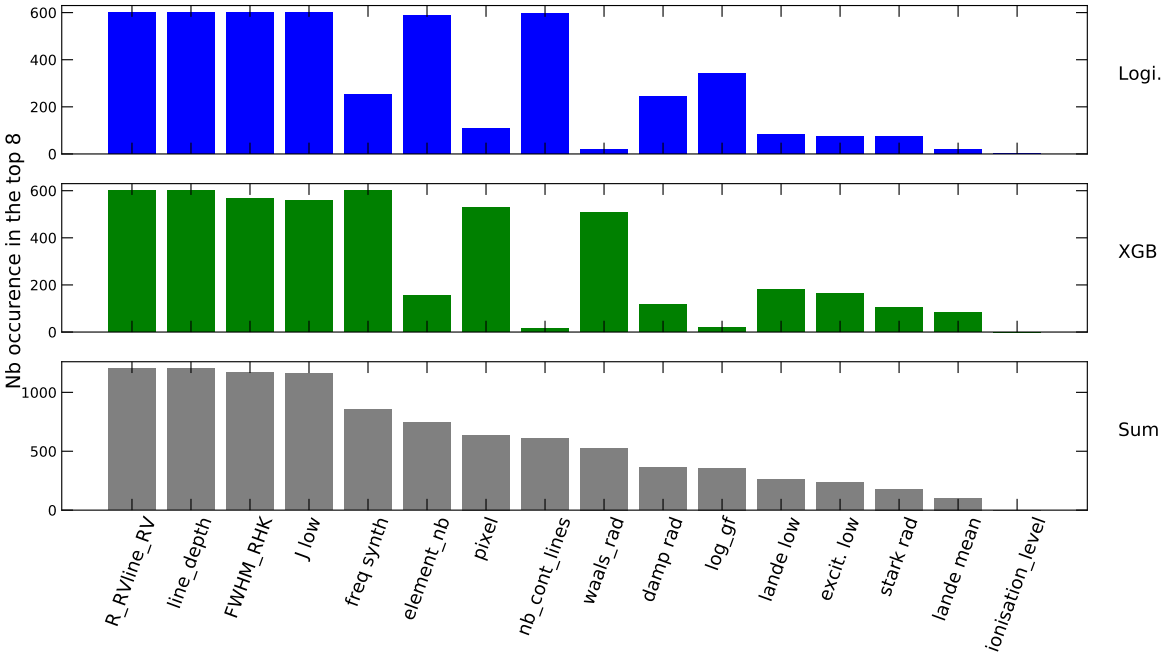

Fig. 33: Occurrence number of the parameters in the top 8 of the 600 simulations for the logistic (top), XGB (middle) and both combined (bottom). For the majority of the simulations, the first 4 parameters are in agreement for both algorithms. But some discrepancies are notable for some parameters underlining the different subprocesses used by the algorithms.

between the occurrence and importance ranking is a sign that the top variables are significant.

Until now, we never mentioned the accuracy of our algorithm. Let us have a quick look on its performance during the 600 test sessions. On Fig.34, we display the precision, recall and F^1 value for the *uncorrelated*, *correlated* and the full selections. The first observation is that for both categories, *correlated* and *uncorrelated*, our predictions are neither randomed nor over-predicted, the score is roughly 97% and no gap is observed between the two recalls. Secondly, we note a slightly better score for the XGB in all categories, that's why for the further analysis we will drop the logistic regression and focus on the XGBoost machine learning algorithm.

We chose to end this section by displaying how the accuracy score change depending on the parameters given. We took the list provided by the top 8 ranking occurrence and suppressed the parameters one by one, first ascending and then descending (progressively from the right and from the left respectively), to show how it impacts the score of the machine learning algorithm (Fig.35). Taking only the $\mathcal{R}$ parameter (best par.), the algorithms reach already an accuracy of 86% which is quite impressive and highlight our previous remark that the frontier in the first parameter plan ($\mathcal{R}$ vs line depth) is almost vertical. By the way, it's not a surprise that $\mathcal{R}$ explains the selection as we remind that the selection are mainly formed using the slope of the correlation between the RV_i and RV_m and it's clear that strong links exist between the $\mathcal{R}$ and the slope itself (the $\mathcal{R}$ is related to the slope and its uncertainty). On the opposite, taking only the worst parameter (the ionisation level), we reach a level close to a random prediction (50%). XGB reaches its highest score as soon as the 5 first parameters with most occurrence are taken.

If the train sample is small compared to the number of parameter, a decrease in accuracy is observed along the way parameters are added, because the machine learning algorithm is lost with the surplus of choice and freedom. But here, given our large sample, such decrease is not observed, and the curve stays flat. Again, XGB algorithm reaches a better score in general compared to a simple logistic regression, motivating our justification to drop the logistic algorithm for the next sections.

Eventually, a relevant comment is to mention that keeping only atomic parameters[4], the accuracy drops at 70%. It's an important remark because for others stars these parameters won't be necessarily derived and thus only the 13 others will be available. By the way, there is a fundamental beauty and deep reasons to understand and distinguish the lines affected by activity only considering atomic physics. Let's also report that when a parameter is deleted, the previous order of importance is not necessarily preserved (see annex to see how the suppression of the $\mathcal{R}$ changed the coefficient power ranking).

[4] removing the $\mathcal{R}$, FWHM_RHK and pixels parameters

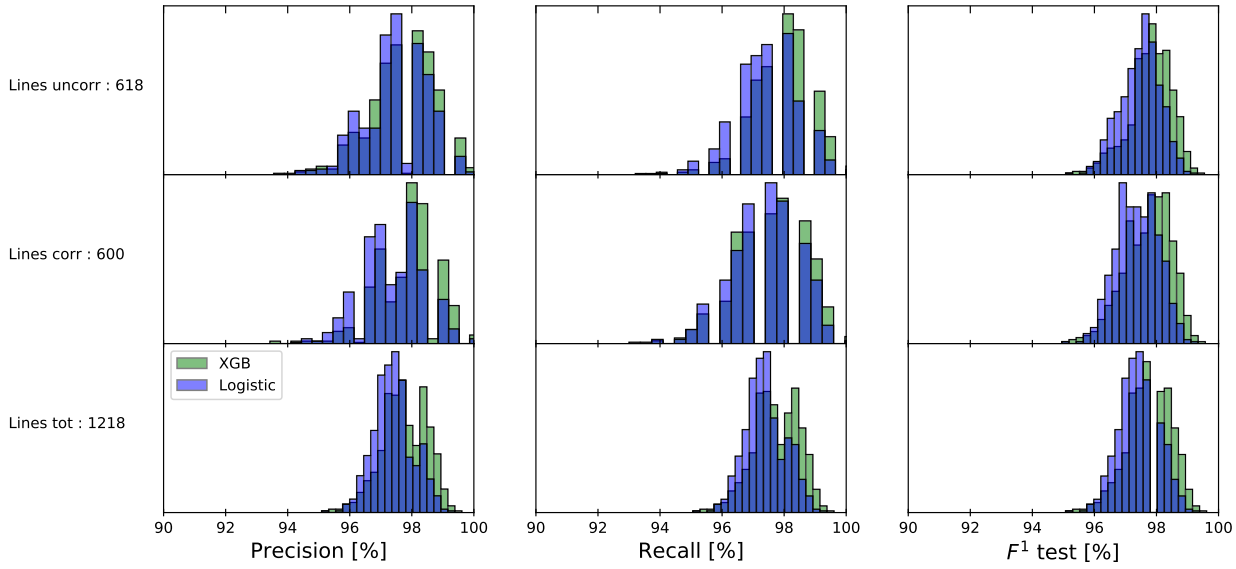

Fig. 34: Summary of the scores obtained for the different classes. **Top** : scores of the *uncorrelated* selection. **Middle** : scores of the *correlated* selection. **Bottom** : global scores. The scores are excellent in most of the simulations (median around 97%), a slight advantage is visible for the XGB algorithm. Precision as well as recall are both excellent, and the fact that no gap is observed between the two recall histograms (top middle and center middle suplots) is a good indicator that no bias toward over prediction is present.

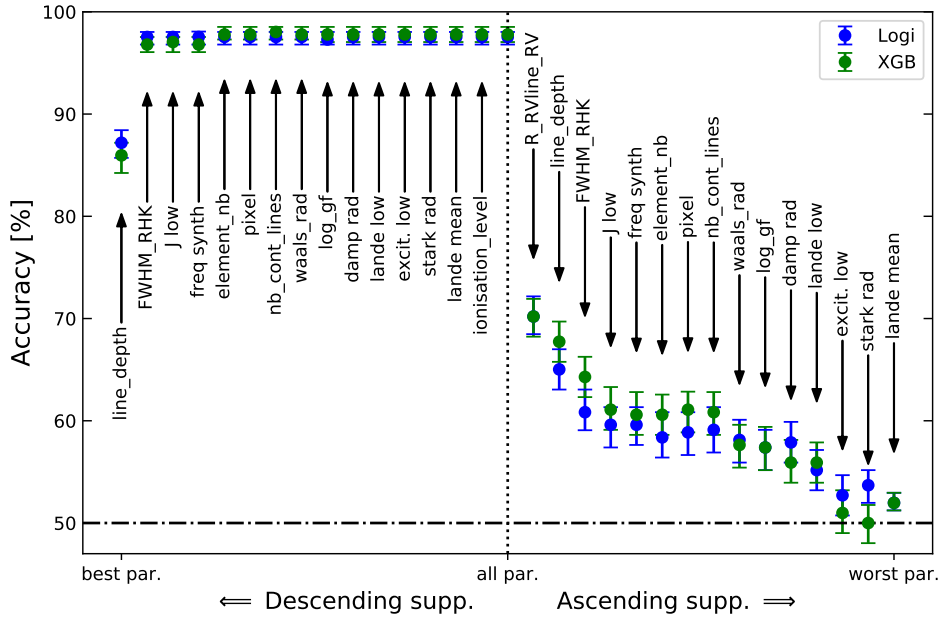

Fig. 35: Accuracy of the simulations depending on the progressive suppression of parameters (the order was taken from the top 8 ranking in Fig.33). The parameter suppressed at each step is written explicitly with an arrow. When all parameters are available, the score reaches an impressive level of 98% of accuracy. The score is already high (86%) only taking into account the best parameter, the $\mathcal{R}$ of each individual line, and the maximum score is reached as soon as the five first parameters are allowed. We outline that the algorithms can still make predictions even with purely atomic physics parameters (dropping the $\mathcal{R}$, FWHM_RHK and the pixel parameters) at a level of 70% of accuracy (not visible on this graphic). And the two algorithms converge after a few suppressions to a level close to a random choice (*dashed line*).

The previous *correlated* and *uncorrelated* selections were giving large and small rms, therefore boosting or mitigating stellar activity, but we never pretend that the thresholds for the slope or the RMS were the best choices. Unfortunately there is a priori no way to guess which threshold of slope, SNR or RMS will minimise the signals' rms except by an optimisation : trying all the imaginable (and wise) combinations. This method is clearly limiting since the size of the grid is exponentially increasing with the number of free parameters. For the *uncorrelated* group, there are 3 free parameters : the minimum slope, the maximum slope and the maximum RMS allowed for a line ; for the *correlated* group : a minimum slope and a minimum SNR. 5 free parameters were already too costly in computation time, that is why we chose to optimise independently the *uncorrelated* and *correlated* selections, generating respectively grids of 3 and 2 dimensions only. Also, let us clarify that instead of suppressing the less precise twin lines (same line appearing on consecutive orders), we kept both of them with the restrictive prescription that if the twins fall in different classes, they are removed and XGBoost will decide of this issue. If XGB also separates twins in different category (the main different parameters between twins being the non-atomic parameters: the $\mathcal{R}$, FWHM_RHK and the pixel), they are removed.

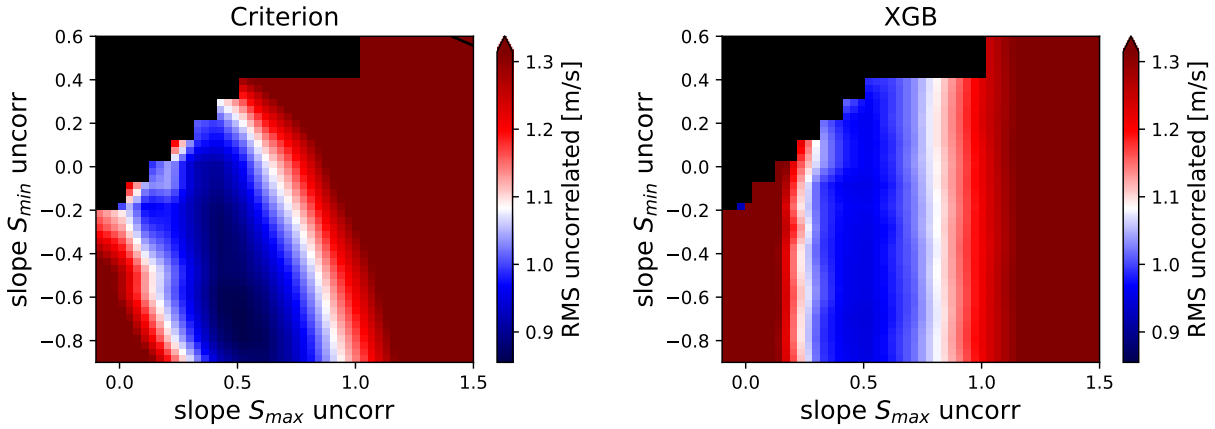

Fig. 36: 2D parameter space projection of the min and max slope (S_{min} vs S_{max}) of the *uncorrelated* optimisation, the last dimension (maximum RMS allowed for a line) should be directed outward the paper but we averaged along this dimension. The color code indicates the rms of the RVs (detrended) obtained with the *uncorrelated* lines (an upper limit at 1.3 m/s have been fixed). Black locations were not in the optimisation grid. **Left** : Uncorrelated rms minimum value from manual selections. **Right** : Same as left with the XGB selections. A well defined and smoothed minimum (86 cm/s) is visible in the *uncorrelated* manual selection indicating that the lines reducing the signal are well located and grouped in the S_{min} vs S_{max} plan. After XGBoost, we observe that the initial S_{min} becomes irrelevant which is expected since the main constrain delimiting the selections are the S_{max} uncorr and S_{min} corr (in Fig.28, it corresponds respectively to the border right of the blue points and border left of the red points, S_{min} uncorr is the left border of blue points and is less relevant).

We display in Fig.36 the result of the parameter space for the *uncorrelated* selection (projection done with an average along the RMS dimension). Because the parameter space appeared not randomly distributed but with a well defined and smoothed minimum (for $S_{min} \sim -0.5$ and $S_{max} \sim 0.5$), we can be confident that the minimum found is the good one and the correlation slope are well established. We note that the XGB algorithm minimum contours are not exactly superposed on the one of the manual selection but still in close vicinity. Also the minimum contours looks spread along the y axis, indicating that the lower limit S_{min} is not helping to reduce the rms.

We display in Fig.37 and Fig.38 the best RV signals obtained with the manual selection and its resulting XGB completion. Let us precise that « the best » does not mean necessary the smallest rms here, since the number of lines was also an important argument in this choice. The fractional weight of the sub-selection (sum of the weight normalised by the total weight) are also indicated to highlight if this selection is initially contributing a lot to the average signal RV_m. Also, because the error is the square root of the weight's inverse, the *uncorrelated* selection with a fractional weight at 26% has error bars roughly $(1/\sqrt{0.26})$ 2 times larger than RV_m. For the case of α Cen B, it is without consequence, since the SNR is so good that the jitter observed is well above the error bars (most of the error bars are not visible whereas they are well present). After XGBoost, we observe than 58% of the average signal can be attributed to active lines, whereas only 33% of the signal is coming from quiet (or less active) lines. The 9% missing are lines with ambiguous behaviour (different selections on the twin lines). It may also look surprising that the fractional weight of the manual *correlated* selection is so low. To understand it let us have a quick look on the Fig.39 which is the parameter space of the observational informations of the RV_i. For this group, the criterion is to select only the highest slope. But the lines exhibiting high amplitude are also the shallowest one and thus the less precise, their weights are in definitive smaller.

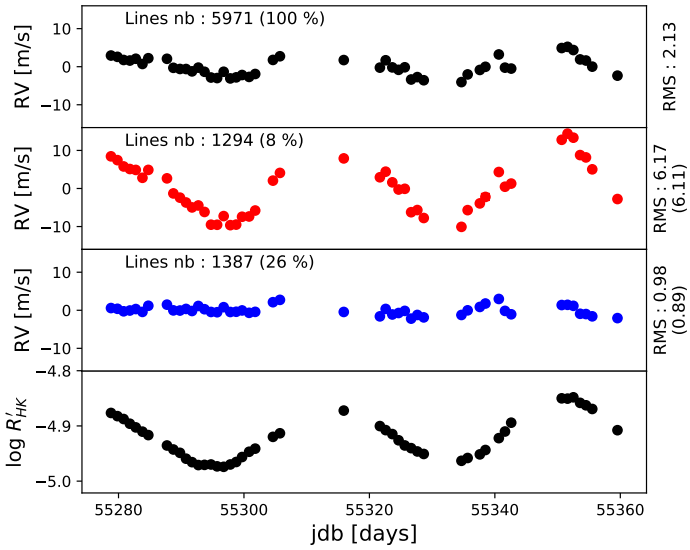

Fig. 37: Best manual selection to reduce the activity contribution in the 2010 RV signal. The average signal using all the lines (***black*** *dots*) is compared to the *correlated* selection (*red dots*) and to the *uncorrelated* one (*blue dots*). In complement to the number of lines, their fractional weight compared to the full sample is displayed in parenthesis. The rms trended and detrended (paranthesis) are also present. The activity index is displayed at the bottom.

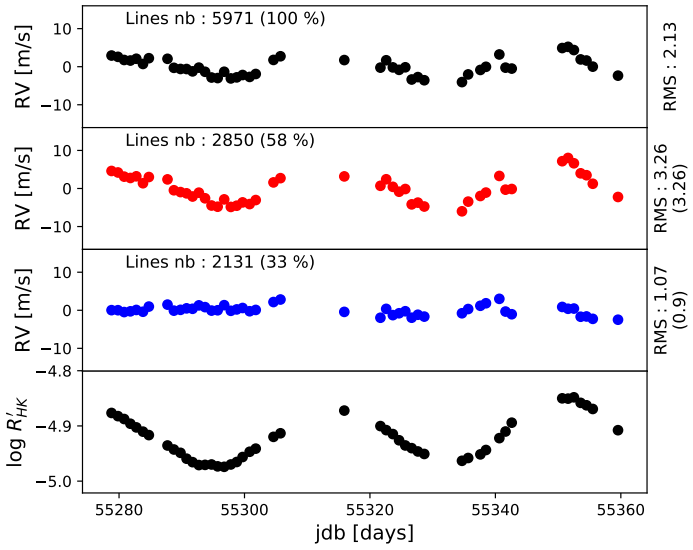

Fig. 38: XGBoost selections from the previous manual groups (Fig.37). The number of lines is substantially larger for a similar uncorrelated rms. The correlated rms decreased as expected but is still 50% higher than the RVs measured on all the lines.

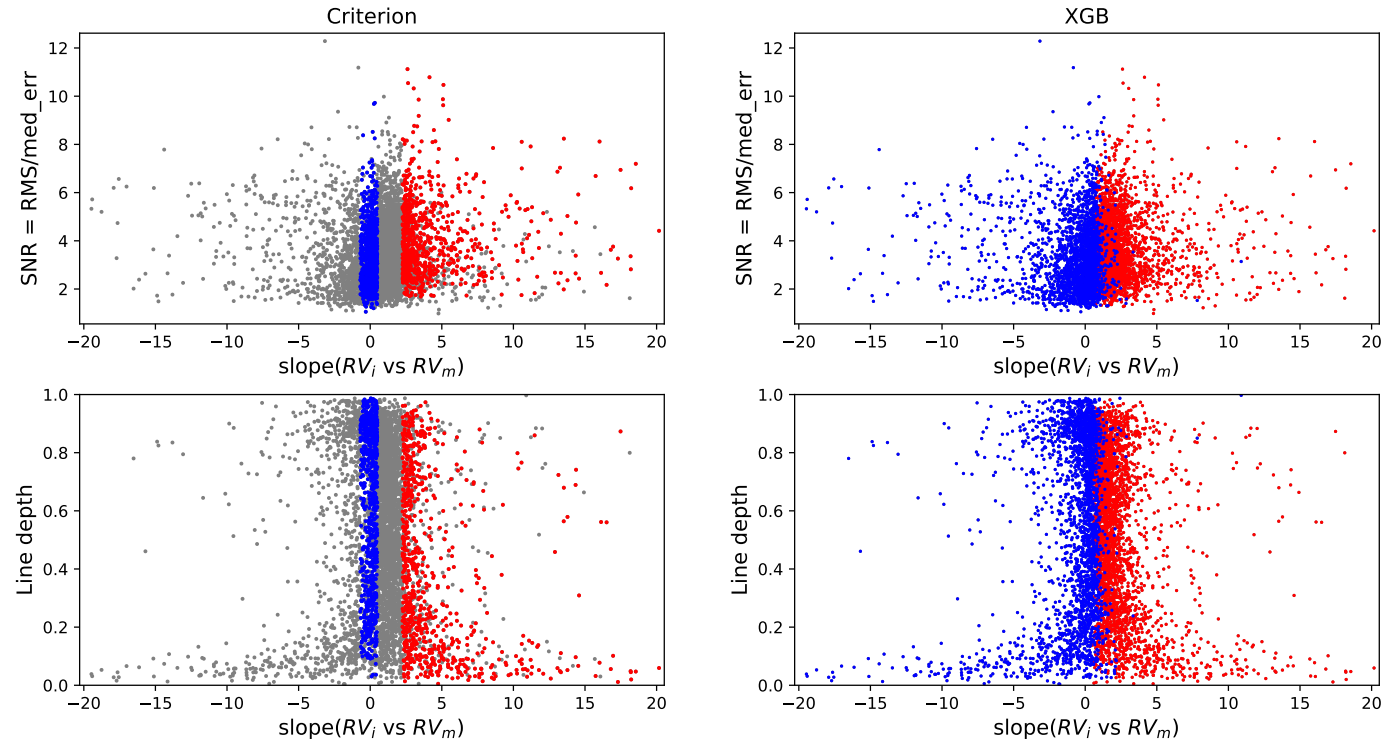

Fig. 39: Observational parameter space. The *correlated* lines (*red dots*) as well as the *uncorrelated* (*blue dots*) are placed surrounded by the other lines of the HARPS K5 catalogue (*gray dots*).**Top** : Parameter space (slope vs $SNR = RMS_i$/median error bars) on which the manual selection is performed. **Bottom** : slope vs line depth. **Left** : Manual input selection. **Right** : XGBoost output selection.

The final purpose of such investigations is to better understand how the activity affects lines and eventually how to use these knowledge on others stars. But before to move on this way, we can wonder if such selections reducing the activity's contribution are time independent, or said differently : are the lines less affected by activity in 2010 capable to mitigate stellar activity of the same star in 2009 or 2011 ? Like for 2010, we derived the RV_i of α Cen B for the year 2009 and 2011. We thus reproduce the exact same analysis that the previous one (selections with the slope of the individual correlation through an optimisation on a grid). The best RV for 2009 and 2011 are displayed below (Fig.40, Fig.41, Fig.42 and Fig.43).

Clearly, the year 2009 is different from the two others, the initial rms was already small (2 times smaller than 2010 and 2011) and there is no "evident" signal of activity as for 2010 where the modulation was visually notable in the $\log R'_{HK}$ as well as in the RVs. Also, the insufficient sampling makes it hard to use the same protocol as in 2010, since less points and no clear feature are two mains obstacles for the reliability of the correlations. For 2011, a clear and long modulation is not visible indicating that the signal may be the composition of several active zones instead of a huge dominant one like in 2010. However, some features are notable as the "W" shape around the $BJD = 55680$ or the "bumps" at 55620 and 55660. The dissimilarity between the $\log R'_{HK}$ index and the RV_m, is either the proof that the signal is composed of several regions or that in this case the photospheric-chromospheric link is poor or that we are affected by instrumental signal. Indeed, there is actuality no necessity for the $\log R'_{HK}$ to perfectly follow the RV signal. First, because the $\log R'_{HK}$ is a measurement of the stellar activity in the chromosphere, whereas the RV is affected by features present in the photosphere and there is not necessarily a one-one relation between the two. Secondly, because if several active regions are present, their mutual contribution produce a complex RV signal. Moreover, α Cen A and B are a binary system and were at their closest distance in the sky in 2015 such that we can wonder if contamination of the spectra by α Cen A is possible for 2011. Finally, these two years, mainly 2009, are less active, which is shown by a smaller $\log R'_{HK}$ index, therefore we expect that the improvement in rms will be less spectacular than for 2010.

Whenever each year is optimised, it is possible to apply these selections on the two others years in order to analyse the compatibility between them. We display the results of the manual selections swapping in Table.4 and the XGB selections swapping in Table.5. An important remark is that the precision of a line is at first order linked with its depth, so because the depth of the lines is not changing between the years, the error on their RV_i (and thus their weights) are almost the same. If the weights were changing the understanding of the situation would be incredibly more difficult because it would be uncertain if a change of the rms would be due to a better selection or to a reorganisation of the lines' weight.

We notice that the selections are only mitigating the year on which they have been optimised. For 2009, we already rises the lack of reliability of the correlations. For 2011, we note that the jumps at $BJD = 55620$ and 56600 (which are the main features and which are the best improvements) does not seem to have a counter part in $\log R'_{HK}$ and therefore, they could be due to instrumental signals. Another problem is the actual unknown presence of planets around α Cen B. For this specific reason, we allowed a second polynomial drift instead of the linear trend of α Cen A. For the clean selection, we tried to suppress only a linear trend but strange variations (longer that the $\sim$ 80 days) were still present providing high rms. The effect could also be induced by the presence of a magnetic cycle. For the year 2011, a clear signal at 12 days is visible in periodogram made on our *uncorraleted* signal, compatible with the period claimed for the hypothetic alpha Centauri Bc (Demory et al. 2015) based on a transit light curve measured in 2013. However, this signal is not found in the data of 2009 and 2010 which have a smaller rms, thus in all likelihood this signal cannot be imputed to a planet. Funnily enough, this period is also the one of Proxima Centauri b (Anglada-Escudé et al. 2016), but α Cen C is definitely to far from α Cen B in the sky to expect contamination from it.

To form our time independent selections, we computed the 2010 and 2011 intersection of lines selections. Because this manipulation provides a really small sample (160 and 216 unique lines with a weight around 2%) in the manual case, we also computed the overlap of the respective two XGB selections, providing a more generous list of lines (1378 and 714 unique lines for weights at 45% and 6%). The intersection has the advantage to remove 50% of the lines randomly predicted by the algorithm. Indeed, if the algorithm is making a random prediction on a line, there is 50% of probability for its prediction to change for the other year.

The lack of overlap between the manual selections could be resolved if lines strongly *correlated* in one of the two year (let us say 2010) are still *correlated* but a bit less on the other year (such lines would in consequence be selected in the manual selection of 2010 but not in 2011 until XGBoost brings them back in the selections). Say again but differently, this observation can be solved dropping the assumption that the importance ranking of lines sensible to activity has to be perfectly preserved with time. Instead of this statement, we can be less restrictive allowing a bit of permeability in the ranking, such that lines in the top can migrate slightly below in the ranking depending on the actives regions.

We also implemented an *exclusive union*. This set is the previous intersection on which we added some particular lines. A line is either *correlated*, *uncorrelated* or *ambiguous*. *Ambiguous* lines are very common in the manual selection and are either lines satisfying no criterion (Table.2) or twin falling in two different classes. After XGB, each line is predicted thus the *ambiguous* lines are only twin predicted in two different classes by XGB. We brought back some of these *ambiguous* lines if during the other year the ambiguity is lifted. Hence, the *correlated* exclusion union is composed of

twin lines *correlated* both years OR *ambiguous* one year but *correlated* the other year. In a similar way, the *uncorrelated* exclusion union is composed of twin lines *uncorrelated* both years OR *ambiguous* one year but *uncorrelated* the other year.

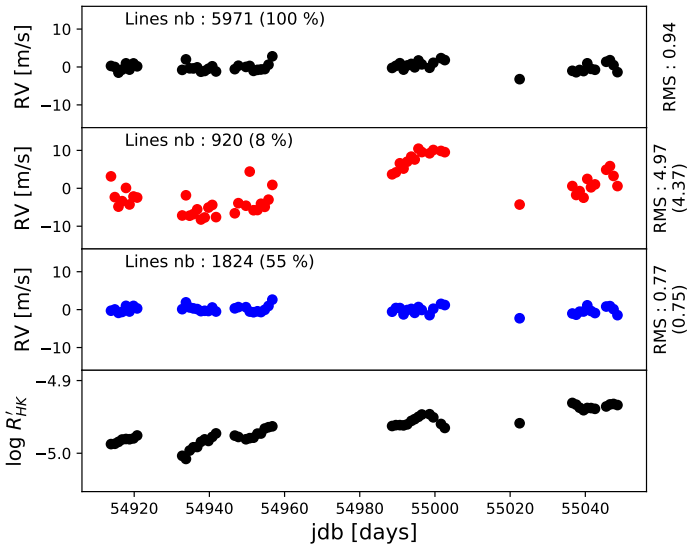

Fig. 40: Same as Fig.37 with the best manual selection to reduce the activity contribution in the 2009 RV signal. The RV initially at 94 cm/s drops to reach a level of 75 cm/s.

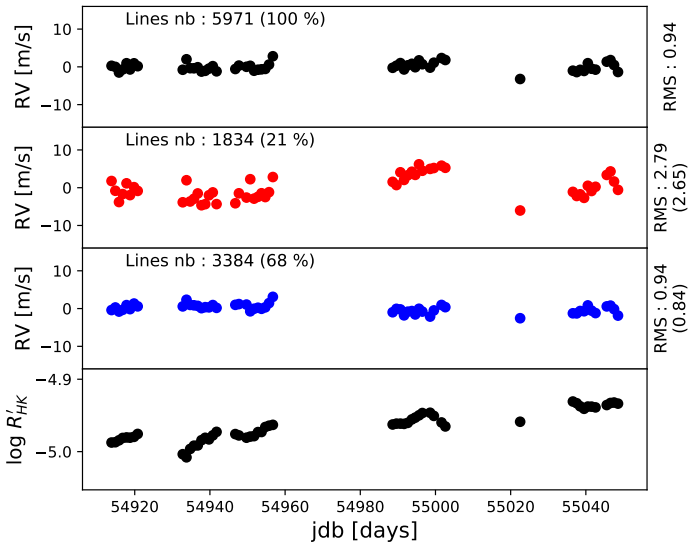

Fig. 41: Same as Fig.38 with the XGBoost selections from the previous manual groups (Fig.40). Opposite to the two others years, majority of lines (68% of the total weight) looks quiet, in agreement with the smaller activity index observed for this year.

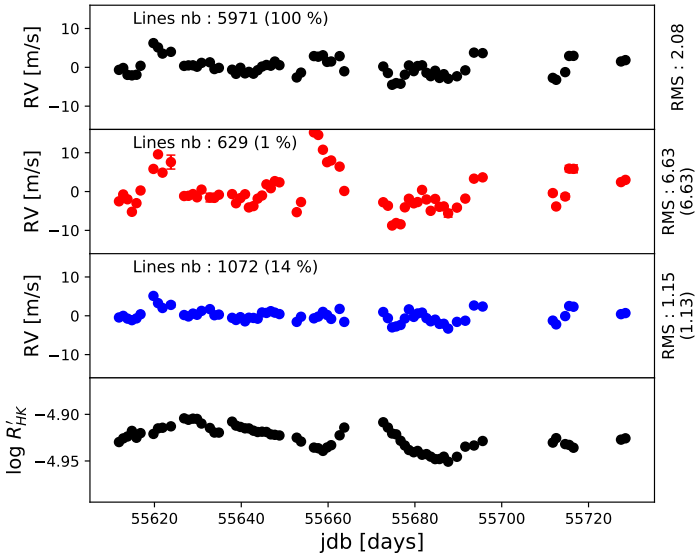

Fig. 42: Same as Fig.37 with the best manual selection to reduce the activity contribution in the 2011 RV signal. The RV initially at 2.08 m/s falls at 1.13 m/s. The best improvement is the reduction of the first two jumps at 55620 and 55660 which are perhaps instrumental.

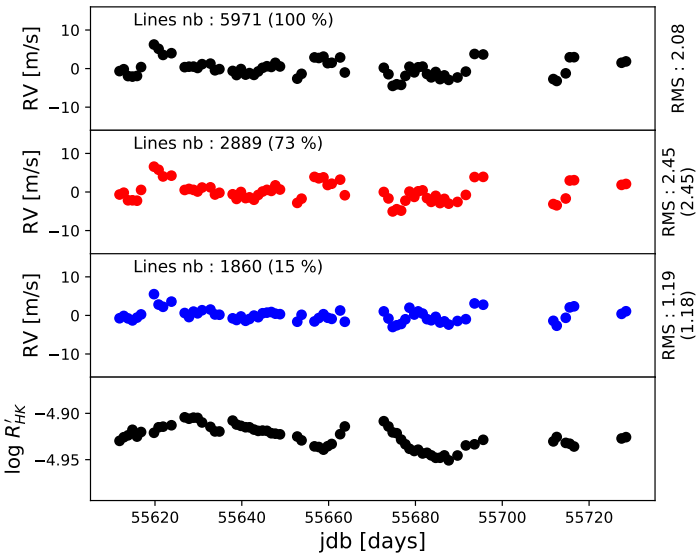

Fig. 43: XGBoost selections from the previous manual groups (Fig.42). We note that the 788 lines added by XGB to the *uncorrelated* selection have almost no weight and thus were lines with a bad precision on them.

Table 4: Summary of the signals' rms for the different best selections of each year. The initial rms of the signal with all the lines is in bold and inside parenthesis. The value for the rms of the *correlated* (upper position) and *uncorrelated* (lower position) selections from the previous graphics are in bold. The number of lines (unique) forming the selections is displayed in the last raw. Applying the best selection of a given year on another one, the rms is always increased/reduced (for corr. and uncorr. respectively) compared to the rms using all the lines (except for the uncorrelated 2010 selection applied on 2009 where a rms of 99 cm/s is reached against 94 cm/s using all the lines). We note a very small intersection between the selections 2010 and 2011, producing a small lines sample correcting very efficiently both 2010 and 2011, but degrading 2009. A more significant sample can be formed with the exclusive union of 2010 and 2011 (last column).

Applied on \ Selection	2009 [m/s]	2010 [m/s]	2011 [m/s]	2010 ∩ 2011 [m/s]	2010 ∪ 2011 [m/s][a]
2009 (**0.94** m/s)	**4.37**	1.67	1.90	3.44	1.57
	0.75	0.99	0.92	1.26	0.94
2010 (**2.13** m/s)	2.94	**6.11**	2.77	7.22	5.43
	2.02	**0.89**	1.95	0.88	1.07
2011 (**2.08** m/s)	2.28	2.44	**6.63**	7.25	3.02
	2.08	2.04	**1.13**	1.14	1.71
Nb. unique lines[b]	705 (8%)	900 (8%)	448 (1%)	160 (1%)	1036 (7%)
	1258 (55%)	951 (26%)	707 (14%)	216 (4%)	1290 (34%)

Notes.
[a] The union of two years is excluding lines in different category between the two years (white intersection of the Venn diagram Fig.45).
[b] In parenthesis is displayed the fractional weight by respect to the full lines available.

Table 5: Same as Table.4 after XGB's run on the previous selections. Conclusion are identical except this time also the 2011 selection degrades the 2009 data. We note a higher intersection between the selections 2010 and 2011 than for the manual selections, producing a substantial sample which corrects efficiently both 2010 and 2011, but still degrades 2009. We point out the rms reached by the union selection in 2010, since 83 cm/s has never been reached in any simulations from the parameter space. We remind that the « exclusive union » performed here allowed to bring back some twin lines that could be classified as different in one year (e.g. *correlated* and *uncorrelated*) but the same in another year (*correlated* and *correlated* or *uncorrelated* and *uncorrelated*).

Applied on \ Selection	2009 [m/s]	2010 [m/s]	2011 [m/s]	2010 ∩ 2011 [m/s]	2010 ∪ 2011 [m/s]
2009 (**0.94** m/s)[a]	**2.65**	1.07	1.06	1.37	1.09
	0.84	0.99	1.00	1.10	1.22
2010 (**2.13** m/s)[a]	2.45	**3.26**	2.15	3.14	3.12
	1.98	**0.90**	1.98	0.95	0.83
2011 (**2.08** m/s)	2.29	2.15	**2.45**	2.45	2.41
	2.04	2.09	**1.18**	1.21	1.23
Nb. unique lines	1534 (21%)	2235 (58%)	2320 (73%)	1378 (45%)	1755 (54%)
	2492 (68%)	1678 (33%)	1474 (15%)	714 (6%)	1016 (10%)

Notes. [a] We got some difficulties to reduce the rms removing only a linear trend since a residual oscillation (magnetic cycle, planet ?) were present. Even with a parabola, some oscillations were persisting.

We could also wonder if such overlap may be due only to chance. As displayed in Fig.44, this is absolutely not possible. We simulated the overlap expected in the case of a random choice for our particular size sample. Let us precise that we took account of the twins for this step since both twins must be predict in the same category to classify a line. There are 5971 lines in the s2d spectrum with non-zero weight. Let us take the example of the overlap occurring in the manual selections (Table.4). We generate a virtual bag containing 5971 papers (total number lines) with number written on them. Alice draw in the bag randomly 1294 papers (the 2010 *correlated* line) without putting them back. When Alice has finished, she keeps the paper and gives the bag to Bob. Bob picks up in the same bag 1387 papers also randomly (the 2010 *uncorrelated* lines). We bring to them a new bag of 5971 papers and ask them to pick again 629 and 1072 papers respectively following the same instructions (corresponding to the year 2011). Alice counts the number of overlap between her two draws. Bob does the same. They note their own number somewhere as well the sum of the two, before to repeat the procedure 10000 times. At the end, the histogram of Alice is drawn (in red), the one of Bob also (in blue) and the histogram of the sum done at each step (in black) on the Fig.44. The fact that Alice keeps her papers secures the possibility for Bob to draw the same number. Mathematically, it assures the disconnection of the two ensembles *correlated* and *uncorrelated* coming from the same year, here from the same bag.

There are two misunderstandings to elude in this story. First, these histograms **are not** showing the average number of lines overlapping randomly in our selections. It would be a huge disaster for us than among our 350 lines uncorrelated overlapping (blue dashed line) roughly 250 could be explained by chance. Again, this graphic is showing that IF there was no link between active/quiet lines in 2010 and those of 2011, the number of overlap would be in average of 250 lines for this particular size of the selections. However, using the counter proposition of the statement we can conclude that the two selections are linked.

The reader may criticise the fact that Alice and Bob draw papers randomly since the criterion to form the groups has nothing to do with random and the same argument for XGB as seen in Fig.34 or Fig.35, but here there is a hidden tricky trick. The fact that Alice and Bob pick paper randomly does not reflect that they choose randomly, it simply ensures us that, mathematically, no information is shared between the first and second bag's papers. In fact, the first random choice performed in the first bag is not necessary. We chose to still perform this first random draw more as a double check. The only obligatory random draw is in the second bag, it allows to assume that even if we have informations (a strategy or something else) for the first bag, this strategy is not useful for the second bag.

Clearly, our results cannot be attributed to chance by more than 8σ for all the overlaps. Better ! Since the space between our overlap number and the random distribution has increased after the XGBoost, it is an unambiguous proof that XGB informations on 2010 and 2011 are sharing some common aspects.

As soon as the reasoning digested, we can perform the same job but on the *mismatching lines*. A mismatching line is a line defined as correlated in 2010 and uncorrelated in 2011 or vice versa (it corresponds to the white intersections in the Venn diagram Fig.45). In total for the manual selection, we founded 245 (69/176) mismatches that we can compare with the number expected from random choice (top right subplot Fig.44). We constate that our mismatching are under the level provided by chance ($>4\sigma$ for all). Actually, the presence of mismatches in the criterion is either the proof that lines affected have changed with time, posing the question if different active regions could not perturb lines in different ways, or that some $\mathcal{R}$ and correlation slopes were spurious. We remind that, in contrast to 2010, 2011 looked more like a superposition of several activity's signals. So we were less enthusiast about the reliability of the correlations.

We are suspecting the lack of a buffer zone to explain these mismatching lines. Let us take an fictive example to illustrate our thoughts. Let a line in 2010 with a perfect sinus shape exactly like the RV_m, at one detail, the phase is shifted by $\pi/2$. This line is neither correlated nor flat. It is a kind of third groups : the dustbin one. Unfortunately, XGB has to put this line somewhere and it has only two solutions : correlated or uncorrelated. That is one issue linked with a binary classification. We displayed such examples for the year 2010 in annexe sectionA.

By the way, a relevant observation is to outline that, suppressing the mismatching lines and gathering 2010 and 2011 XGB selections (explicitly : forming the exclusive union of the two selections), the 2010 rms reaches a value of 83 cm/s never obtained by any simulation. A Venn diagram[5] (Fig.45) as been drawn to help the reader with the different sub-selections performed in the two tables.

Furthermore, until now, the binary classification restricted us to considered that a line is either *correlated* or *uncorrelated*, but visually we noted that in 2010 some lines (a very few number) exhibit RV *anticorrelated* with the average one (an example is displayed in Fig.A.1). Because an anticorrelation is still a correlation, it means that these lines are still affected by the activity in a sense. We performed a visual inspection on the XGB exclusive union selection to flag and remove them. This operation suppressed 41 lines, such that the final uncorrelated sample is formed by 975 lines (weight still 10%). On the 2010 data the rms of the signal falls at its minimum value of 77 cm/s against 1.22 m/s for 2011.

[5] A classical Venn diagram of 4 subsets cannot be draw with circles. However, our situation is less general since two pairs of circle are not crossing. Nevertheless, the reader may keep in mind that our figure is just a stylish mind support instead on the true one.

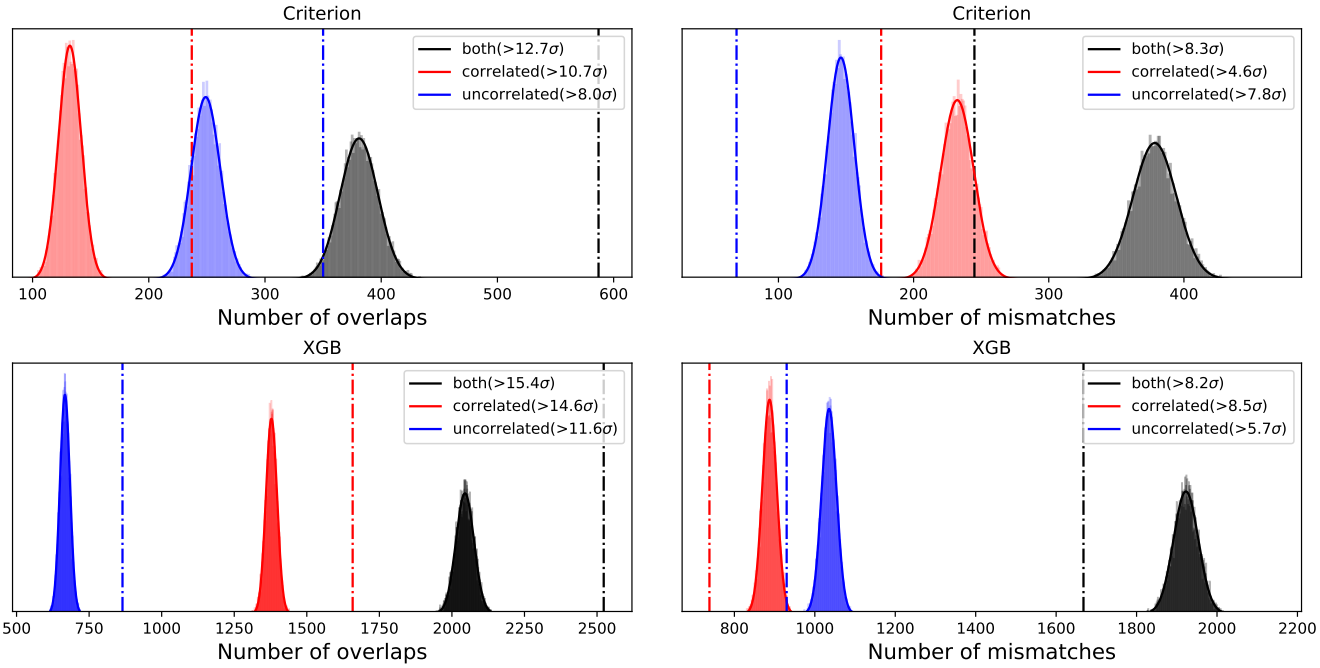

Fig. 44: Density probability distributions of the number of overlaps (left) and mismatches (right) occurring between 2010 and 2011, if the lines' "identities" were absolutely not related (overlap or mismatch will in consequence be a question of chance). A Gaussian function has been fitted over the 10'000 simulations. The number of overlaps between the *correlated* selections (*red*) as well as between the *uncorrelated* selections (*blue*) are displayed. The number of overlaps occurring regardless the selections is in black. **Top** : For the manual selections. There is roughly more than 8σ separating the probability mean random distributions to our overlap numbers (*dotted-dashed lines*). Thus, the overlaps observed are not due to chance. The number of mismatch from its part is slightly below the level of chance by more than 4σ. **Bottom** : From the XGB selections. There is even more distance between our overlaps number and the distribution ($> 10\sigma$), confirming that the XGB methodology is not random and that 2010 and 2011 lines share some underlying properties. The number of mismatches increases in absolute but slower than the number expected from random choice ($> 5\sigma$).

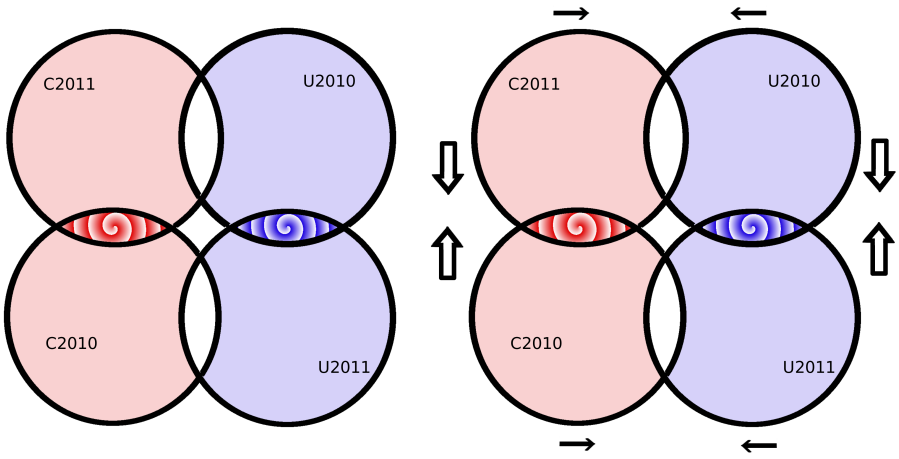

Fig. 45: Venn diagram for the visualisation of the selections (area are not proportional to the true size sample). There is no intersection between the *correlated* (C) and *uncorrelated* (U) selections of the same year. The intersection selections (spirals) are the overlap between the two years, the exclusive union (transparent colors without white + spirals) the gathering of the two years but excluding the diverging predictions. The exclusive union allows to bring back twin lines with an ambiguous behaviour during one of the year, if this uncertainty is lifted during the other year. The centered square contains lines were the behaviour is ambiguous during both years. **Right** : Translation of the Fig.44 in Venn space. XGB increases the size of the intersections well beyond the level expected by random choices (large arrows). It also increases the number of mismatches, but relatively to the random level, the number of mismatch is smaller than with the manual criterion, that is why we draw it with small arrows.

6. Investigations on the Sun

6.1. The 8 planetary RV contribution

The Sun is a peculiar star, not compared to other stars but because of its proximity with the Earth. Such proximity bring some advantages, like a good SNR of the spectra even if the exposures are short or the fact that the Sun surface can be spatially resolved with a small ground telescope (obviously the Sun if already resolved by eye). By the way and almost by definition, the Sun is the only star which can be observed during the day. Another distinction comes from the fact that in contrast with other stars, the Earth is gravitationally bounded to the Sun. It introduces some differences in the classical formulae. The RV semi-amplitude K induced by a body of mass m_p on another body of mass m_* is given by :

$$K^3 = \frac{2\pi G}{P} \cdot \frac{(m_p \sin(i))^3}{(m_* + m_p)^2} \cdot \frac{1}{(1-e^2)^{3/2}} \quad , \tag{4}$$

where e is the eccentricity, P the period of revolution and i the inclination of the system. For the case of the Sun (assuming $e \simeq 0$, $i \simeq 90°$ and $m_\odot \gg m_p$), the formula can be reduced for units of m/s :

$$K \simeq 0.0891 \cdot \left(\frac{m_p}{M_\oplus}\right) \cdot \left(\frac{P}{years}\right)^{-1/3} \qquad \text{or} \qquad K \simeq 28.3 \cdot \left(\frac{m_p}{M_♃}\right) \cdot \left(\frac{P}{years}\right)^{-1/3} \tag{5}$$

The idea to move on the Sun is not only for ease of reference, it is also because, for our own star, the planetary system is so well known that we know precisely what we MUST see in the RV. The table below (Table.6) displays the semi-amplitude of the RV for our own solar system (for the end of the work, we will use indistinctly the word semi-amplitude or amplitude). Until now, there is no difference compared with exoplanetary system. The difference is hidden in the fact that the period of the RV signal measured from the Earth is not the true revolution period (P_{rev}) but the synodic one (P_{syn}). Indeed, as an example, Jupiter will produce a signal of roughly 12 m/s in amplitude and 400 days in period (instead of its 12 years revolution's period). This is because the Earth is also turning around the Sun and much faster than Jupiter. The RV signal is modulated by the periodicity of Jupiter **and** by the position of Earth on its orbit. The synodic period can be simply computed from the difference of the angular velocities (it is a difference because all planets are rotating in a codirection) :

$$\omega_{syn} = \omega_{planet} - \omega_\oplus$$

So expressed with the periods ($\omega \propto 1/P$):

$$P_{syn} = \frac{P_\oplus \cdot P_{rev}}{|P_\oplus - P_{rev}|} \quad , \tag{6}$$

with $P_\oplus$ the Earth's sidereal period (365.25 days). The absolute value allows to neglect if the planet is situated inwards or outwards the position of the Earth's orbit. Without the absolute value, since the planets outside the Earth's orbit have a longer period of revolution (3^{rd} Kepler law), the synodic period would be negative, translating the fact that the ω_{syn} angular vector is pointing in the opposite direction that the initial ω_{planet}. We see that the formula is not defined if the planet is rotating with the same period than the Earth, since the synodic angular velocity ω_{syn} is null. It means that we cannot find the Earth signal when looking to the Sun from Earth.

Table 6: RV semi-amplitude (K) and synodic period (P_{syn}) for the planets of the solar system (ordered by ease of detection). The semi-amplitude has been computed with the exact formula Eq.6.1 instead of the approximated one.

Planet	Mass [$M_\oplus$]	P_{rev} [years]	K [m/s]	P_{syn} [days]
Jupiter ♃ :	318	11.85	12.49	399
Saturn ♄	95.3	29.42	2.76	378
Uranus ♅ :	14.5	83.75	0.30	370
Neptune ♆ :	17.3	163.7	0.28	368
Venus ♀ :	0.816	0.62	0.09	584
Mercury ☿ :	0.055	0.24	0.01	116
Mars ♂ :	0.108	1.88	0.01	780
Earth ♁ :	1	1	0.09	∞

Eventually, limits can be taken in the two extremes cases of «hyper-close» or «hyper-far» planets. Taking the logarithm of the previous equation :

$$ln(P_{syn}) = ln(P_{\oplus}) + ln(P_{rev}) - ln|P_{\oplus} - P_{rev}|$$

Thus, for $P_{rev} \ll P_{\oplus}$ like the case of Mercury, $P_{syn} \approx P_{rev}$. In the opposite, if $P_{rev} \gg P_{\oplus}$ like for our four gas giants, $P_{syn} \approx P_{\oplus}$.

To sum up, we do not have to deal with the Earth's signal since the Earth cannot see its own RV vector when looking to the Sun. The Earth only modifies the period measured for the others planets. We simulated in Fig.46 the last 20 years effect of the giants on the Sun (assuming no ellipticity and no inclination of the system). We display in Fig.47 the RV of the Sun measured « from another » exoplanet (situated toward the Vernal point[6]) and from the Earth. For humans, the presence of Saturn induces an oscillation in the Jupiter's signal of 2.8 m/s on time scale given by the least common multiple (LCM) of their two synodic periods : LCM(399,378) $\sim$ 19.7 years. The two icy giants contribute even less in this deviation. So on period of a few years the presence of the three other giants is a linear modulation.

After the signal of the 4 giants, the next signal to find is Venus at 10 cm/s, such exquisite precision level is difficult to reach but another interesting observation is the fact that Venus should produce an unambiguous signal at a period different by 200 days of the giants ones. However, at such levels not only the stellar activity but the instrumental noise of HARPS begins to limit the detection. Because Mars and Mercury are far to small in the contribution (<1 cm/s) there is no hope to detect them.

For the Sun, since the position of all the planets is precisely known, we can perform a change of frame from the barycentric to the Sun's center. This operation has been done and has for effect to remove all the planets contribution.

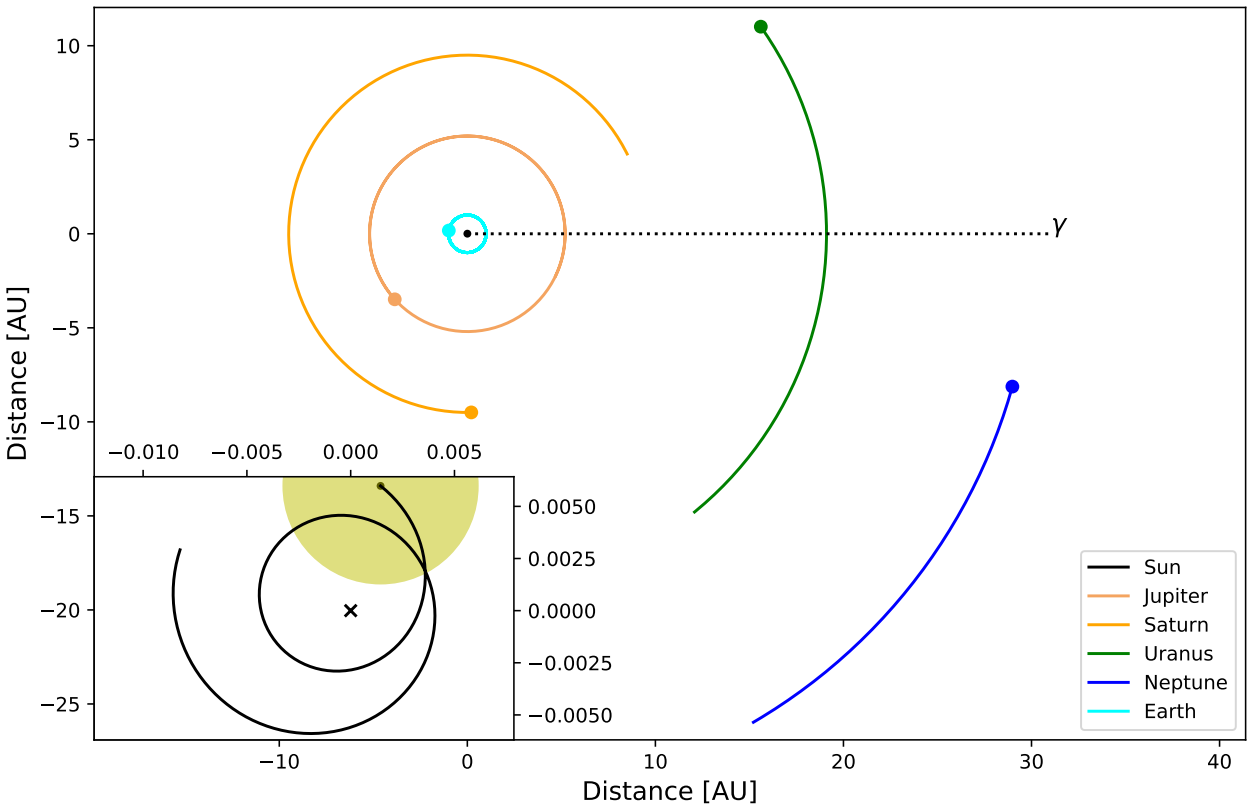

Fig. 46: Simulation of the impact of the giants planets on the Sun in a simple vision where all the planets are in a perfect plan with circular orbits around the barycenter, without interaction between them. **Outer plot** : Evolution of the giants planet on their orbits during the last 20 years (*solid lines*) ended by their current position (*dots*) March 16th, 2017. The vernal γ direction (*dashed line*) is indicated. **Inner plot** : Sun displacement around the barycenter (*cross*) induced by the giants, the units are also in AU. The barycenter is laying just outside the Sun's surface (*yellow area*). The most important contribution is coming from Jupiter and Saturn.

[6] In the solar vicinity, the K2.5 dwarf HD4628 (also called *96 Psc* or *Gliese 33*) situated 7.5 pc away would fulfill roughly this condition

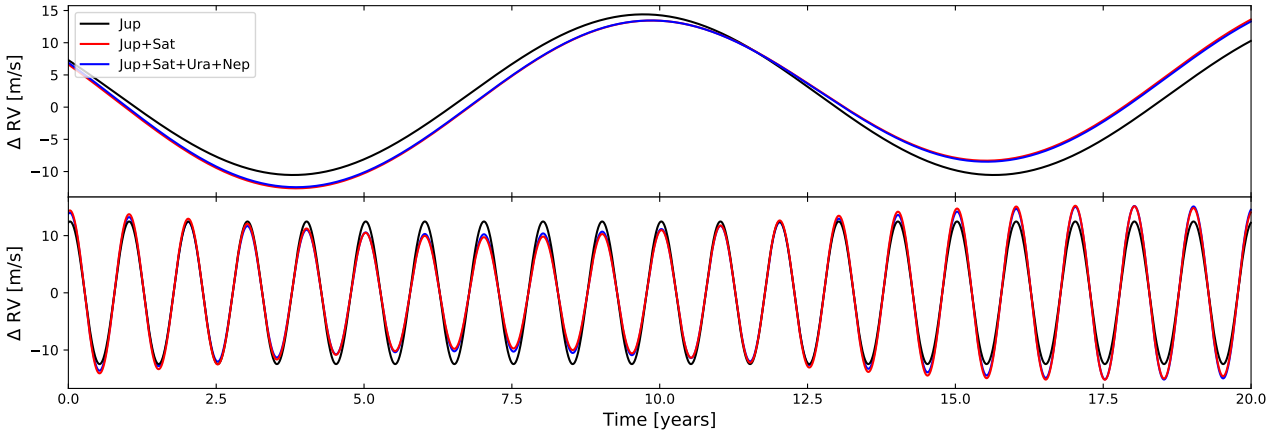

Fig. 47: RV signal of the Sun taking into account Jupiter (***black*** *solid curve*), Jupiter and Saturn (*red solid curve*) and the four giants (*blue solid curve*). **Top** : RV of the Sun measured by an exosystem located toward the Vernal direction (for instance around *Gliese 33*). By definition of the Vernal point, the exoplanet see our system edge on. Jupiter induces a 12.5 m/s signal with a period of 11.9 years. Saturn's presence is only visible due to a small phase shift compared to Jupiter's signal. **Bottom** : RV of the Sun measured from the Earth. The RV amplitude of the giants are the same that for an external observer however the period measured are the synodic ones taking into account the Earth revolution. Saturn modulates the Jupiter's signal between 9.7 and 15.3 m/s (12.5 ± 2.8 m/s) on period of 20 years (least common multiple between their synodic periods).

6.2. Reduction of the Sun's contamination

We performed the same work as for α Cen B using the Sun data in 2016 after the change of rest frame to the Sun center. Unfortunately for us, at this time, the Sun was almost at the minimum of the 24^{th} Schwabe cycle which is itself one of the least active cycle since one century. By the way, the 24^{th} cycle is still the field of research for its particular inactive level (Li et al. 2011). When the Sun is active, variation of a few m/s should be observed (Borgniet et al. 2015), but in our case the rms is small for a value of 1.22 m/s. Nevertheless, clear modulations are visible. We optimised respectively, the two first bumps (first shadow window in Fig.48) as well as the two others coming later (second shadow window). The $\log R'_{HK}$ signal seems to match with the second bump of the first window and with the W shape of the second window. Results for the best selections are shown in Table.7 and Table.8 following the same schema than for α Cen B.

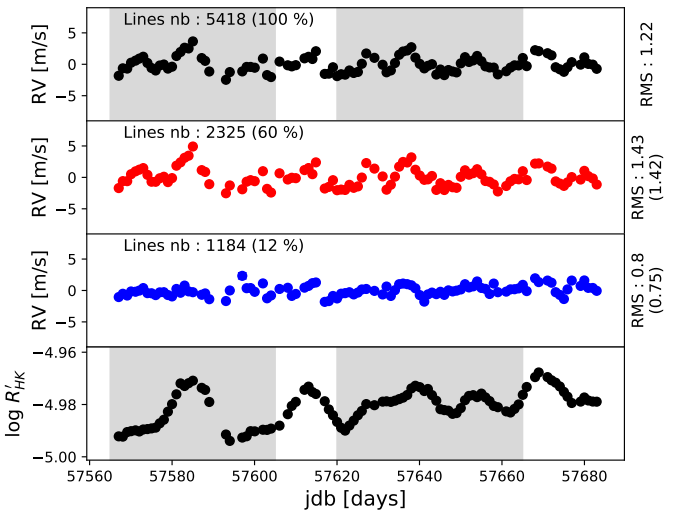

Fig. 48: RV using the XGB exclusive union formed by the selections in the first and second windows (*shaded area*). The observations start in July 2016 and end in October 2016. Some bumps in RV are not present in the $\log R'_{HK}$ signal.

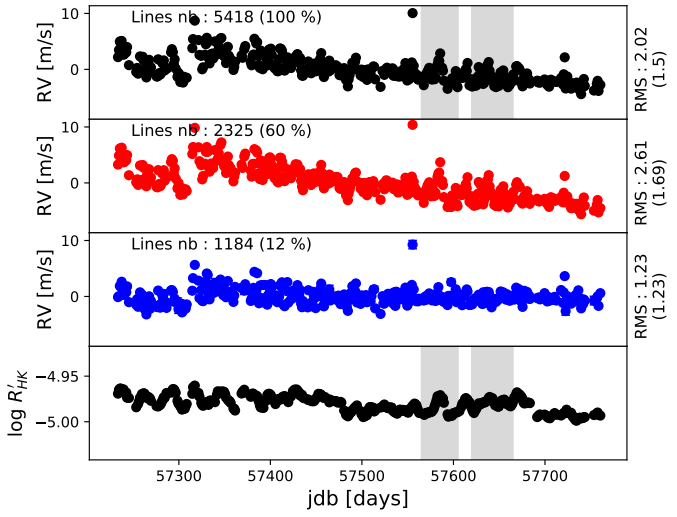

Fig. 49: Same as Fig.48 for observations starting in August 2015 and finishing in January 2017. On so long timescales, the solar magnetic cycle is expected to be visible. Because the RV are free of planetary signals and that the trend is affected by our selections, it confirms that the linear drift is induced by the magnetic cycle.

Table 7: Same as Table.4 for the Sun radial velocities. Two different regions have been optimised (shadowed windows of top subplot Fig.48).

Selection / Applied on	Sun 1^{st} bumps [m/s]	Sun 2^{nd} bumps [m/s]	$1^{st} \cap 2^{nd}$ [m/s]	$1^{st} \cup 2^{nd}$ [m/s][a]
Sun 1^{st} window	**5.38**	1.64	4.58	1.78
(**1.31** m/s)	**0.72**	1.35	0.74	1.12
Sun 2^{nd} window	2.70	**2.32**	4.18	2.14
(**1.19** m/s)	1.06	**0.82**	0.85	0.85
Sun 2016	4.14	2.04	4.73	1.92
(**1.22** m/s)	1.03	1.09	0.91	1.00
Nb. unique lines	385(1%)	919(8%)	184(1%)	1088(8%)
	322(13%)	546(22%)	113(5%)	723(29%)

Notes. [a] The union of the two selections is again an exclusive union.

Table 8: RMS resulting from the XGB of the previous selections for the Sun. Two distinct temporal regions of the 2016 data have been optimised. The selections formed have been applied on the 100 days and 1.5 years measurements. For the latter, the linear trend induces by the magnetic cycle is also indicated.

Selection / Applied on	Sun 1^{st} window [m/s]	Sun 2^{nd} window [m/s]	$1^{st} \cap 2^{nd}$ [m/s]	$1^{st} \cup 2^{nd}$ [m/s]
Sun 1^{st} window	**1.59**	1.39	1.55	1.52
(**1.31** m/s)	**0.54**	1.32	0.67	0.62
Sun 2^{nd} window	1.26	**1.49**	1.46	1.42
(**1.19** m/s)	0.99	**0.75**	0.66	0.63
Sun 2016	1.36	1.42	1.45	1.42
(**1.22** m/s)	0.93	1.06	0.78	0.75
Sun 2015 - 2017 (detrended)[a]	1.68	1.67	1.76	1.69
(**1.50** [m/s])	1.34	1.37	1.31	1.23
Sun 2015 - 2017 (lin.trend)[a]	-4.69 ± 0.24	-4.79 ± 0.24	-5.41 ± 0.26	-5.21 ± 0.22
($\mathbf{-3.56 \pm 0.20}$ [m/s/year])	-0.71 ± 0.18	-1.32 ± 0.19	0.42 ± 0.17	-0.40 ± 0.15
Nb. unique lines	2028(71%)	2167(58%)	1265(48%)	1707(60%)
	1393(15%)	1264(25%)	597(7%)	837(12%)

Notes. [a] Because there is sufficiently long data and we are sure that the linear drift is not induced by a planet we wrote for this special case the rms of the detrend signal and the slope of the linear trend.

As for α Cen B, we remark that the selections correct only slightly (or degrade) the others temporal regions after selections swapping. Performing as before the intersection and the exclusive union (Fig.48), this strategy allows to form a sample working everywhere. At best, the signal on the $\sim$ 100 days reaches a rms of 75 cm/s which means an improvement of 40% by respect to the initial rms (122 cm/s) using all the lines. Again, this choice of selections among all the possibility given by our grid optimisation was the one balancing the best the rms reduction and the weight of the *uncorrelated* selection.

We noted that some RV modulations are not visible in the $\log R'_{HK}$ signal itself, likely because the RV signal is composed of many active regions like α Cen B in 2011. Another observation that we would like to emphasise is that, after visual inspection, some lines are showing the 57625 BJD peak (first bump of the second window) but not the followings (Fig.50). On the opposite, some lines were showing the 'W' modulation around 57650 (also visible in the $\log R'_{HK}$ index) but not the previous peak (Fig.51). Although, the jitter makes difficult a reliable interpretation.

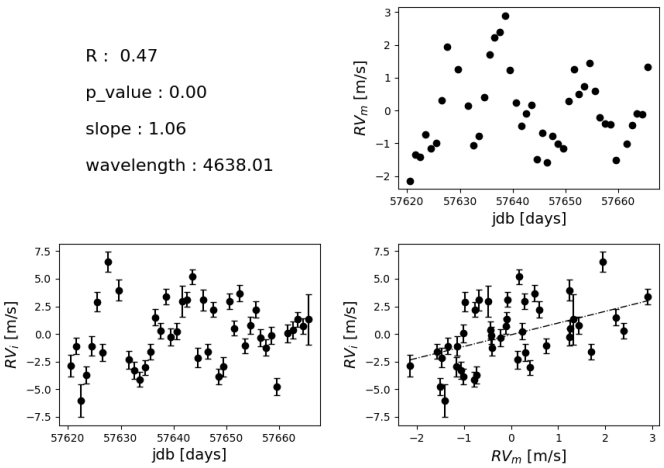

Fig. 50: Same as Fig.8. A line in the Sun showing the first peak but not the following W modulation.

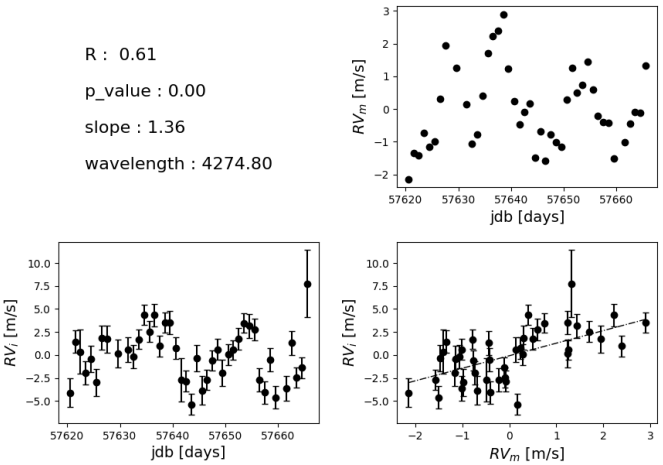

Fig. 51: A line in the Sun showing the W modulation but not the previous peak.

The Sun provides a new ground of investigation about stellar contamination thanks to the excellent and long covering time of the observations. The perfect knowledge of the solar system lifts unknown and allow us to investigate the second problematic contamination coming from long magnetic cycle variation. For α Cen B, such investigations cannot be performed since the observations were not continuous and that such variations could be explained by unknown planets. We applied the XGB union selection on the Sun's RV data which start in August 2015 and finish in January 2017 (Fig.49). After removing one outlier (BJD=55453), we performed a weighted linear regression to compute the slope S of the drift. The different trends are :

$$S_m = -3.56 \pm 0.20 \left[\frac{m/s}{year}\right] \qquad S_{corr} = -5.21 \pm 0.22 \left[\frac{m/s}{year}\right] \qquad S_{uncorr} = -0.40 \pm 0.15 \left[\frac{m/s}{year}\right]$$

The selections of lines based on local consideration are also mitigating long magnetic fluctuations. Because spots and faculae are driven by the magnetic field of the stars, there is actually good reason to think that the same lines are affected by long magnetic variation. Actually, the drift of 33 cm/s is really impressive since we are now very close to the detection of Earth-like planet at 10 cm/s. We could criticise that the jitter is still of 1.2 m/s, but if the jitter is purely Gaussian (white noise) this is not an obstacle for the detection of planets with smaller amplitude, only most numerous measurements and on longer timescale are needed.

7. Correlation of the activity signal with the $\log R'_{HK}$ index

Our method allows to form a « *synthetic activity signal* » (SAS) produced by subtracting the *correlated* signal from the *uncorrelated* one (we used again the XGB union selections). It is possible to show that the SAS is free from planetary or binary RV signal. Let's write explicitly the equations. The *correlated* RV (RV_{corr}) is the weighted average ($<>_i$) of the *correlated* lines, using the equation (3.1) for the individual radial velocities :

$$RV_{corr} =< RV_i >_i \equiv \frac{\sum_i w_i \cdot (RV_p + RV_{a,i})}{\sum_i w_i} = \frac{RV_p \sum_i w_i + \sum_i w_i \cdot RV_{a,i}}{\sum_i w_i} = RV_p \frac{\sum_i w_i}{\sum_i w_i} + \frac{\sum_i w_i \cdot RV_{a,i}}{\sum_i w_i} = RV_p + < RV_{a1,i} >_i$$

Same for the *uncorrelated* group :

$$RV_{uncorr} = RV_p + < RV_{a2,i} >_i$$

With $RV_{a2,i} < RV_{a1,i}$ if the group *uncorrelated* has been well formed. Since all lines are affected in the same way by the planetary signal (we thus explicitly neglected the i indices on the RV_p term), it implies that the SAS becomes free of planetary (or binary) signal :

$$SAS \equiv RV_{corr} - RV_{uncorr} =< RV_{a1,i} >_i - < RV_{a2,i} >_i$$

The SAS is at the end noisier than each individual selection, but the important point is that it only contains activity's effects or instrumental ones (which should be small compared to the activity). Next step consists to investigate if the SAS is closely or not related to the most important activity indicator : the $\log R'_{HK}$ index. In Fig.52, we display the SAS for the 3 years of α Cen B and for the 1.5 years data of the Sun in Fig.53. The first comment is to mention that the SAS and the $\log R'_{HK}$ are well correlated for the two stars. About α Cen B, the most apparent feature is the incredible similarity between the SAS and the $\log R'_{HK}$ in 2010 ($\mathcal{R} = 0.91$). Again this is because, at this time, an exceptional huge region was dominating the star activity. This is not the case for 2009 and 2011, since for the first year almost no activity is visible and the sampling is quite irregular, nevertheless, the global trend and the $BJD = 55000$ bump are notable ($\mathcal{R} = 0.75$). While the third year is likely composed of many regions providing a SAS more complex with a long trend still coherent with the $\log R'_{HK}$ ($\mathcal{R} = 0.56$). The three slope are giving an average value of 68 ± 3. Let us remark that the same values are not necessarily expected between the slopes, since they depend on the size and latitude of the regions (projection considerations) which changes as a function of activity level. Moreover, we point out that the SAS is actually only inducing a redshifted contamination. This is expected invoking the effect of convective blueshift inhibition which dominates the spots' effect for slow rotator observed almost pole-on (Dumusque et al. 2014) like α Cen B.

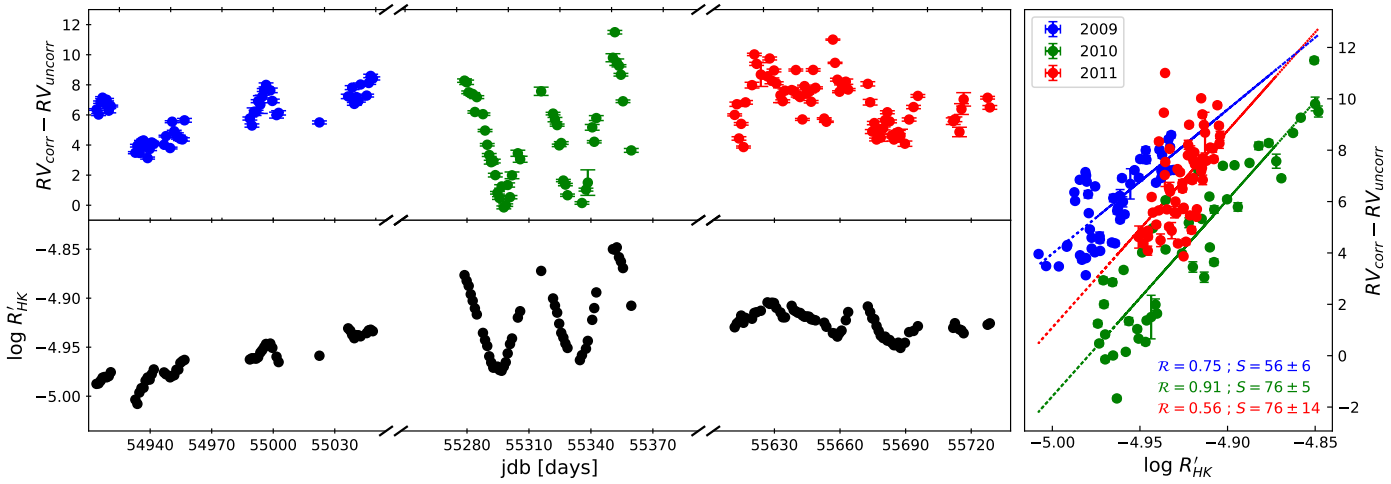

Fig. 52: Comparison between the « synthetic activity's signal» (SAS) produced taking the difference between the *correlated* and *uncorrelated* RV and the $\log R'_{HK}$ index for the α Cen B star. Broken axis have been used on the time axis to avoid the empty space caused by the absence of observations between the years the scale is however the same for the three years. The global trend as well as small features are quite similar providing a nice correlation between the two signals ($\mathcal{R} > 0.56$). **Upper panels** : SAS as a function of time for the 2009 (blue), 2010 (green) and 2011 (red) data. **Lower panels** : $\log R'_{HK}$ activity level for the same time observation than the upper panels. **Right panel** : Correlation graphic.

For the Sun, we can remark that the $\log R'_{HK}$ undergoes some jumps induced by instrumental issue on HARPS-N caused by ghost contamination on the CCD (private communication). This contamination increases the flux in the core of the Ca II H and K line, therefore increasing artificially the $\log R'_{HK}$. To suppress this effect, the detector is thus regularly heated, producing the brutal diminutation emphasised by small arrows on the Fig.53. We corrected manually this effect with the assumption that the ghost contamination is evolving linearly and that the true value is the one after the heating

in order to reestablish the continuity. The three warm up are indicated by the arrows and delimit 4 «segments». Let us remark that actually the first warm up ($BJD = 57308$) vanishes the offset in the $\log R'_{HK}$ but well visible in the RV, likely caused by the presence of 4 active regions on the visible hemisphere of the Sun (Fig.E.2). We can only compute the linear correction for the 2^{nd} and 3^{rd} segments (8.6e-5 and 7.7e-5 $days^{-1}$) for which the two boundaries warm up are known. For the 1^{st} and 4^{th} segments, we took for the correction the average of the two previous linear corrections assuming when necessary than the warm up are roughly equidistant in time. After such manipulations, the $\mathcal{R}$ reaches 0.83 and the slope 227 ± 8, the residual is about 1.2 m/s.

As for α Cen B, we emphasise than the SAS is mainly composed of a redshifted contribution expected by faculae. The general trend is negative since at this period the Sun was on the declining part of its magnetic cycle. After $BJD = 57700$ we note that the SAS becomes slightly negative (-1.3 m/s) which is quiet unexpected since faculae should only produce a redshifted contribution and spot alternatively a redshifted contamination directly followed by a blueshifted contribution. However the RV_i were determined relatively to spectra establish when the solar activity was slightly more active ($\log R'_{HK} = -4.95$) explaining perhaps the blueshift. Identical modulations are found in the SAS and the $\log R'_{HK}$ for all observations preceding $BJD = 57400$, when the Sun was more active. Around $BJD = 57590$, we note a huge redshift (5 m/s) followed by an important blueshift (-5 m/s). The redshift can be assimilated to the particular extended and numerous active regions at this time : from AR2562 until AR2568. In particular, the regions AR2565 and AR2567 gave birth to large sunspots in SOHO/SDO imagings (the latter was even being born on the visible side). We do not remarked visual counterpart to explain this blueshift, but since there is only 1 measurement, its reliability is not ensured.

It is more difficult to compare the slopes between the stars since several factors affect the values : the inclination angle, the rotational period, the size of the regions, the metallicity, the spectral class. The fact that our SASs are so related with the activity level is the proof that our method is mitigating the stellar activity. By the way, the fact that a perfect matching is not found is an indicator that, except for some rare cases, it is not possible to use the $\log R'_{HK}$ signal to correct the RV since the RV effect induced by some features are not necessarily correlated with the $\log R'_{HK}$. We remind that the index is not providing direct information about the photosphere. Moreover, because the $\log R'_{HK}$ is based on the observations of a few and close-frequency lines, this index is more easily sensitive to instrumental issues as ghost.

Some authors tried to use the $\log R'_{HK}$ and the BIS of the CCF as proxies of the activity to correct the RV_m signal with for instance Gaussian Process (Rajpaul et al. 2015), but because each line is affected differently, this solution is too much reductive. Jones et al. (2017) improved the Rajpaul's work, generating their own proxies of activity using principal component analysis (PCA) method. Their work provide impressive results, even if their simulations were applied on a "simple" case of a singular huge spot (1%) for an equator-on star (RV peak-to-peak of ~ 14 m/s), thus no conclusion are raised in more complex and realistic case where several active regions of divers types or sizes are evolving. Haywood et al. (2016) are indicating that the $\log R'_{HK}$ is maybe not the most precise activity indicator. The best indicator would be the magnetic flux computed taking into account high spatial resolution magnetograms and Dopplergrams. However, since the authors computed the $\log R'_{HK}$ using solar spectra reflected on Vesta, some scattering in blue may have occurred, providing an unreliable $\log R'_{HK}$. This method of Zeeman-Doppler imaging was used by Donati et al. (2017) to reduce the activity around a T Tauri star (rms $\sim$ 650 m/s), successfully discovering a hot Jupiter ($K \sim 70$ m/s). However, Zeeman-Doppler imaging is extremely difficult to measure in the case of slow rotators like the Sun.

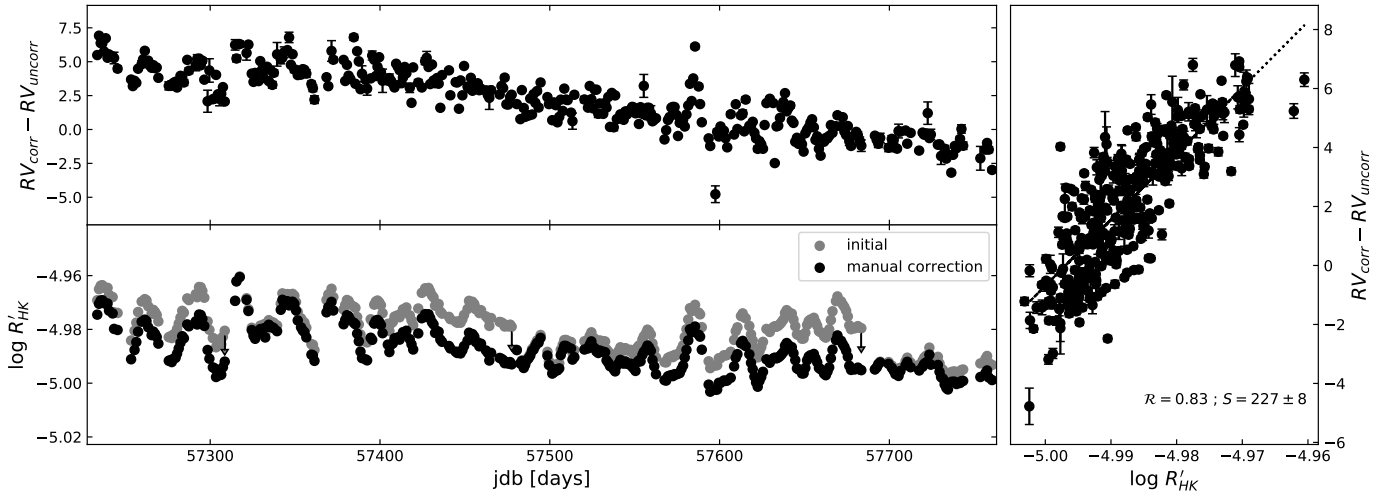

Fig. 53: Same as Fig.52 for the Sun. The global trend as well as local features are similar providing a good correlation between the two signals ($\mathcal{R} = 0.83$). **Upper panels** : SAS as a function of time. **Lower panels** : $\log R'_{HK}$ activity level for the same time observation than the upper panels. The initial index (*gray dots*) is displayed as well as its manual corrected version (***black** dots*). The warm up are indicated (*arrows*). **Right panel** : Correlation graphic.

8. Spectral type dependency of the selections

The *uncorrelated* sub-selections formed before are the lines the less affected by the activity for respectively α Cen B in 2010 and 2011 and for the Sun in 2016. A legitime interrogation is to wonder if these selections are exchangeable or present some overlap one with respect to the other. The Sun and α Cen B are different spectral type stars, respectively a G2 and K1 main sequence stars, clearly, these classes are not the closest ones possible but neither the more distant. The next investigation is to find out if the activity affects the same lines regardless their spectral type, we do not expect a perfect match for many reasons. First the spectral classification is precisely a classification based on spectral features (width, depth, lines density). It means that some lines visible in the Sun's spectrum will be missing in the red dwarf's spectrum and conversely. Our samples being already small, the absence of lines is a problem. Moreover, α Cen B is more metallic and colder than the Sun. It means, in one hand, that lines in α Cen B are deeper than for the Sun (see Fig.54), since the line depth is the main atomic parameter which can predict if a line is strongly affected by activity, the selections for *correlated* and *uncorrelated* lines could be different. In the other hand, the convective envelop extends deeper than for the Sun, such that we cannot neglect the possibility for their magnetic field dynamic to be different, producing distinct activity's signatures. As an example, we mention the result of the simulations of Beeck et al. (2013a) and Beeck et al. (2013b) about the convection features and the effect on the granulation. At the optical surface level, the convective flows are roughly twice slower for a K0 star than for a G2 spectral type, the upflows cover a less extended area and the granulation is four times smaller in size.

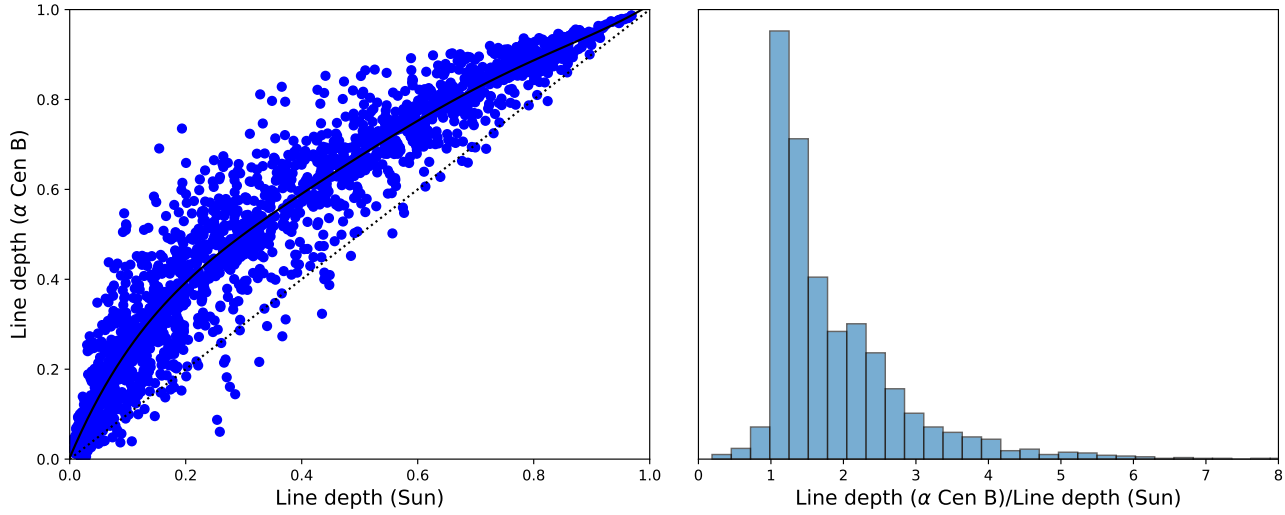

Fig. 54: Comparison (for lines present in both stellar spectrum) of the line depth observed for the Sun and α Cen B. Most of the α Cen B lines are deeper than in the Sun. A polynomial fit (*solid line*) has been performed to highlight this aspect in comparison with the same depth level (*dotted line*). In average, the depth is ~ 1.7 times higher for α Cen B.

In Table.9, we present the rms of the selections swapping between the two stars for the most important XGB selections established before. The top left and bottom right part of the table (diagonal by block) were already presented in previous tables. We can see that both selections are actually less efficient to reduce the contamination of the other star activity, but the improvement is still notable except for the solar selections on α Cen B 2011. Such results were expected and could be explained by the remark announced before but also for others reasons. Firstly, for the second window of the Sun, we noted that some lines were showing RV variation for some features and not for others. Secondly, we noted that the selections performed on different temporal features were also less efficient to mitigate the stellar activity on another temporal range. We remind that this is the initial reason explaining why an intersection and an exclusive union have been performed. It raises the eventuality that activity's signatures are more complex that we initially thought such that depending on the nature of the active regions (temperature, size, phase of the magnetic cycle) different lines can be affected.

We would like to focus more deeply on this point since both stars are maybe not affected by the same features. The inclination angle is playing for instance a key role in the balance between spots and faculae from a photometric aspect. For a question of contrast, faculae are more visible in the limbs that at the disk center whereas the reverse is true for spots. As a consequence, the balance tilts more and more to the faculae contribution when the inclination angle is close to 0 (Radick 2001). α Cen B is a slow rotator likely see almost pole-on from the Earth, this configuration implies that the main contamination is coming from faculae and not spots. On the flip side, the Sun is observed equator-on and rotates slightly faster (the Sun is however still a slow rotator with $v_{rot} \sin i$ about 2 km/s) such that the contribution of spots could play a more important role than for α Cen B. Moreover, because the active regions were presumably at different latitudes and the two stars at different phases of their magnetic cycle, it is possible for their underlying magnetic field to be different which imprints a «personal footprint» on the spectral lines.

Table 9: Investigations about the XGB selections swapping between the K1 dwarf α Cen B and our own G2 star. The upper left and lower right part of the table were already written in previous tables, the 2009 selection of α Cen B has been removed because of its low reliability. A new exclusion union is performed between the two stars providing a new and last selection (column $\biguplus$). As before, the initial rms using all the lines is displayed in bold and parenthesis in the left column. An improvement between $15-30\%$ by respect to the initial rms is colorised in dark green, whereas an improvement $> 30\%$ is displayed in blue. The units are in m/s or m/s/year.

Applied on \ Selection	α Cen B 2010	α Cen B 2011	2010 $\cup$ 2011	Sun 1^{st} window	Sun 2^{nd} window	$1^{st} \cup 2^{nd}$	$\biguplus$
α Cen B 2009	1.07	1.06	1.09	1.00	0.97	0.98	1.03
(**0.94** [m/s])	0.99	1.00	1.22	0.93	0.95	0.91	1.17
α Cen B 2010	3.26	2.15	3.12	2.24	2.26	2.31	2.72
(**2.13** [m/s])	**0.90**	1.98	**0.83**	1.99	1.90	**1.77**	**0.94**
α Cen B 2011	2.15	2.45	2.41	2.08	2.10	2.10	2.30
(**2.08** [m/s])	2.09	**1.18**	**1.23**	2.11	2.04	2.08	**1.49**
Sun 1^{st} window	1.42	1.34	1.43	1.59	1.39	1.52	1.50
(**1.31** [m/s])	1.17	1.23	**1.07**	**0.54**	1.32	**0.62**	**0.69**
Sun 2^{nd} window	1.30	1.22	1.29	1.26	1.49	1.42	1.38
(**1.19** [m/s])	1.07	1.16	**0.99**	**0.99**	**0.75**	**0.63**	**0.68**
Sun 2016	1.34	1.25	1.33	1.36	1.42	1.42	1.39
(**1.22** [m/s])	1.10	1.18	**1.02**	**0.93**	1.06	**0.75**	**0.76**
Sun 2015 - 2017 (detrended)	1.61	1.54	1.61	1.68	1.67	1.69	1.65
(**1.50** [m/s])	1.34	1.38	1.30	1.34	1.37	**1.23**	**1.12**
Sun 2015 - 2017 (lin.trend)	-4.76 ± 0.22	-4.28 ± 0.22	-5.04 ± 0.23	-4.69 ± 0.24	-4.79 ± 0.24	-5.23 ± 0.24	-5.07 ± 0.24
(**-3.56 $\pm$ 0.20** [m/s/year])	**-2.10 ± 0.18**	**-1.89 ± 0.16**	**-1.45 ± 0.18**	**-0.71 ± 0.18**	**-1.32 ± 0.19**	**-0.40 ± 0.15**	**0.18 ± 0.14**
Weights	58%/46%	73%/61%	54%/42%	52%/71%	45%/58%	41%/60%	61%/66%
α Cen B / Sun[a]	33%/30%	15%/12%	10%/8%	19%/15%	24%/25%	12%/12%	11%/10%

Notes. [a] Fractional weights computed for the 120 days of the Sun, the values can slightly change for the 2 years data.

A last exclusive union ⊎ is done between the two stars to mitigate the activity for both stars. This selection is quite efficient on every observation windows and for the first time the long trend magnetic variation of the Sun is reversed (slope positive) and barely null taking into account the uncertainty (0.18 ± 0.14 m/s/year). Also, we note that the rms of the linear detrended signal of the ⊎ selection reach a level of 1.12 m/s which is better than the solar selection at 1.22 m/s.

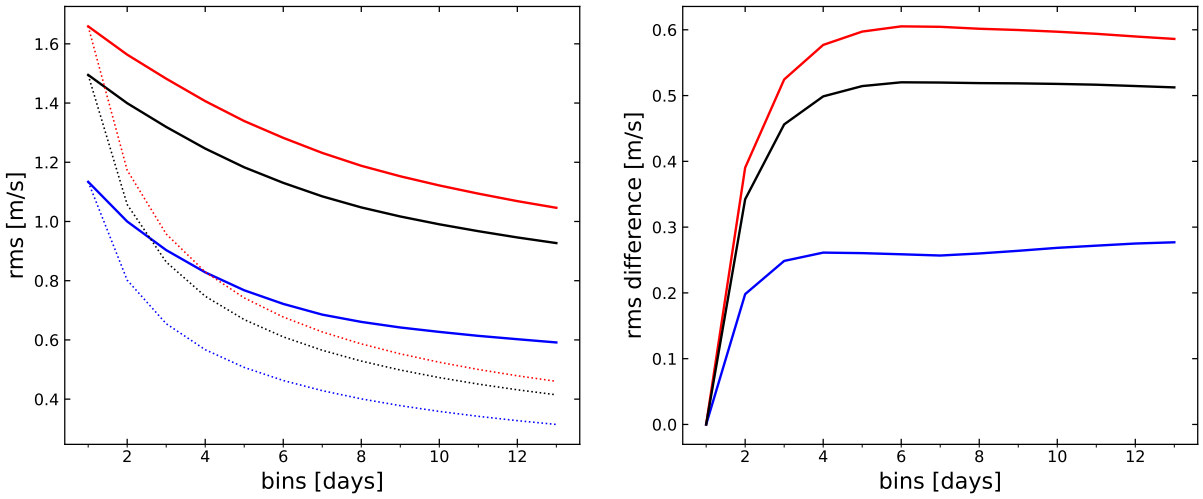

Fig. 55: **Left :** Rms of the detrended RV signals (1.5 years data of the Sun) as a function of the number of days used for the binning (with the exclusive union selections ⊎ formed by the Sun and α Cen B). The *uncorrelated* signal (*blue curve*) is roughly 35 cm/s better than the CCF signal (***black** curve*) and 55 cm/s better than the *correlated* signal (*red curve*). The decrease expected from Gaussian uncertainties ($\propto 1/\sqrt{N}$) is also displayed (*dotted curves*). **Right :** Difference between the Gaussian noise and our signal rms. A red noise is still present under 4 days and the level of red noise is twice lower for the RV_{uncorr} than for the RV_m or the RV_{corr}.

During this work, we never reached lower rms values than $\sim$ 70 cm/s for the 1-day binned RV data except on short period like for the first 40 days of the Sun observations (Table.8) where a value of $\sim$ 50 cm/s was reached. Because the SNR is very good for both stars, this deviation is not coming from the photon noise but must be induced by other stellar mechanism (a good suspect would be the supergranulation) or instrumental limits precision. The precision of HARPS is estimated at about 50 cm/s on short timescale (shorter than 1 night), and at 1 m/s for longer dispersed measurements even thought the lack of stable sources make difficult such threshold establishment (Mayor et al. 2003). This instrumental hypothesis will be elude very soon when the successor of HARPS, the ESPRESSO spectrograph, will provide its first results, since its precision should be of $\sim$ 10 cm/s (Pepe et al. 2014). For the Sun, supergranulation pattern is found to have a typical overturn time of a few days (Toomre et al. 2001; Del Moro et al. 2004; Rieutord & Rincon 2010), such that a binning over several nights can reduce this effect. The granulation lifetime is about a few minutes (Del Moro 2004; Beeck et al. 2013b) and thus should be already mitigated. In Fig.55, we display how the rms is falling down when more days are used in the binning. In particular, the remaining red noise is visible below 4 days and its level is lower in our *uncorrelated* signal than in the RV_m or RV_{corr}. A problematic aspect is that since only 10% of the spectral information is used, the error bars would be increased by a factor 3. However, the transition from a red noise (jitter variating with the observations because of activity) to a white noise (jitter time-independent) is an advantage that could balance the loss of precision.

As a closure for this section, we will compare the exclusive unions from the XGB selections for the two stars, since these selections are often efficient in general and are sufficiently populated. Because a line with a twin in one star could be deprived of twin in the other stars, the less precise twin is suppressed for the analysis. We have 4419 unique lines with a weight not null for α Cen B, for the Sun this amount is of 4053. We possess the RV of both of them for 3465 lines (meaning that already $\sim$ 1000 lines (respectively 600) have disappeared from α Cen B (respectively the Sun) which corresponds to roughly 25% of the selection (respectively 15%). For α Cen B, among this 3465 lines, 1439 were correlated and 741 uncorrelated, against 1438 and 725 for our star. The two stars share 680 lines correlated and 218 lines uncorrelated in common. For a number of mismatch of 491 lines. We remark that the two selections are not perfectly matching which was expected for the reasons announced previously. As we did before, we can compared these numbers with a random choice meaning that the selections between the two stars share no information at all. All the distributions from random classification (overlaps as well as mismatches) are distant of more than 4σ from our values, such that a significant amount of lines are found affected in both stars. This result is also supported by the fact that the selections, without being excessively performant, are affecting the rms in the good direction. In definitive, even though the two selections cannot be exchanged to correct efficiently the RV of the other stars, some links may exist in the lines affected between the two stars.

9. Improvement on the minimum mass detection

Until now, we only measured the improvement to mitigate the stellar activity through the rms computation performed on the RV signal, but such indicator does not provide any detailed information on the planetary detection improvements. As an example, with a 1 m/s rms RV signal, it is still possible to detect amplitude of 0.5 m/s if the noise is purely white and the observation sufficiently long. So is our *uncorrelated* signal able to detect smaller planet ? Does it depend on the period of the planet ? Are the orbital parameters more constrained ? To resolve all these questions, we work with the 120 days RV data of the Sun in 2016. We simulate the presence of planetary signal by injecting different periods P, semi-amplitudes K and phases ϕ in the RV signal. We focused on the special configuration of circular orbits, hence the RV signals injected are purely sinusoidal.

The procedure is the following : the planetary signal is injected in each spectral line at the beginning. We then compute the 1-day binned RV signal *uncorrelated* and the average one, before to remove the linear trends. The signals are recentered in time for numerical stability before to be given as input for a Lomb-Scargle periodogram analysis (oversampling of 10). We look and save all the periods at local power maxima and check their significance with a p - value of 1%, computed by bootstrap. Then, a χ^2 minimisation by the Powell algorithm is performed to find the best parameters (semi-amplitude, period and phase), using the period given by highest power peak in the periodogram as initial guess. Once the best fit is founded, a Monte Carlo Markov-Chain (MCMC) with flat prior is performed with the best fit as a starting point.

For the Sun, spots are known to appear mainly at low and intermediate heliographic latitudes between ~ -30 and $30°$. According to the *butterfly diagram*, at beginning of the Schwabe cycle spots appear in average around $\pm 30°$, whereas, at the end of the cycle, they are situated close to the equator. At these latitudes, the average differential rotational period is about ~ 25.5 days, which corresponds once the synodic modification is taken into account about ~ 27.5 days. Small departures from this value are expected if not only one but many spots are visible on the stellar surface, if faculae are dispersed on large surface and considering that many regions have a shorter lifetime than the rotational period and are evolving. In Fig.56, we display the periodogram of the signals without planets, corresponding, thus, to the periodogram of the black and blue detrended signals in Fig.48. We clearly note the peak induced by the activity at P_{rot} even though only its first harmonic corresponding around $P_{rot}/2$ is higher than the 1% p - value. In our *uncorrelated* signal, the first harmonic is just sufficient to exceed the 1% p - value level. In Fig.57, we draw, as a visual example of the improvement, the results of a 3.4 $M_\oplus$ planet injection with a period of 41 days and an amplitude of 65 cm/s. Whereas the planet is detected with our lines selections, the planet is not found using all the lines.

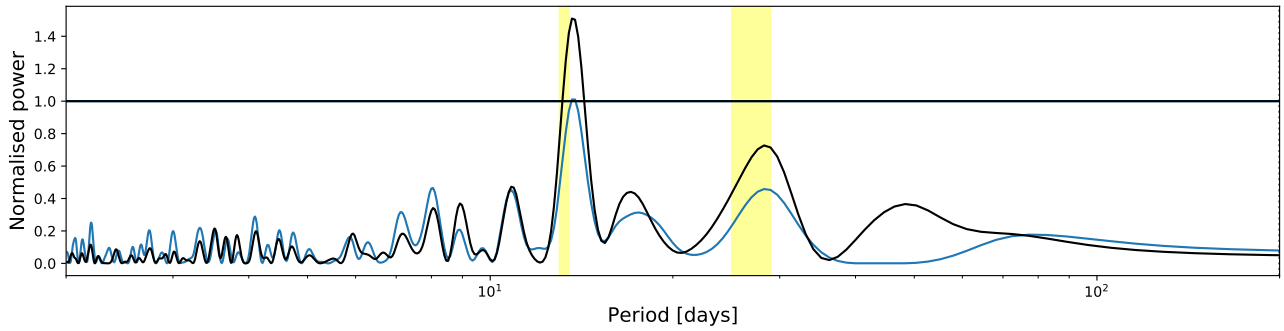

Fig. 56: Lomb-Scargle periodograms performed using all the lines (***black*** *curve*) and using our *uncorrelated* lines (*blue curve*). The 1% p - value level is displayed (*horizontal lines*) and the power of both periodograms is normalised by their own respective 1% p - value level. The expected zone of solar activity contamination (synodic) at P_{rot} and $P_{rot}/2$, the first harmonic, are indicated by yellow stripes, where the width is to into account the differential rotation of the Sun.

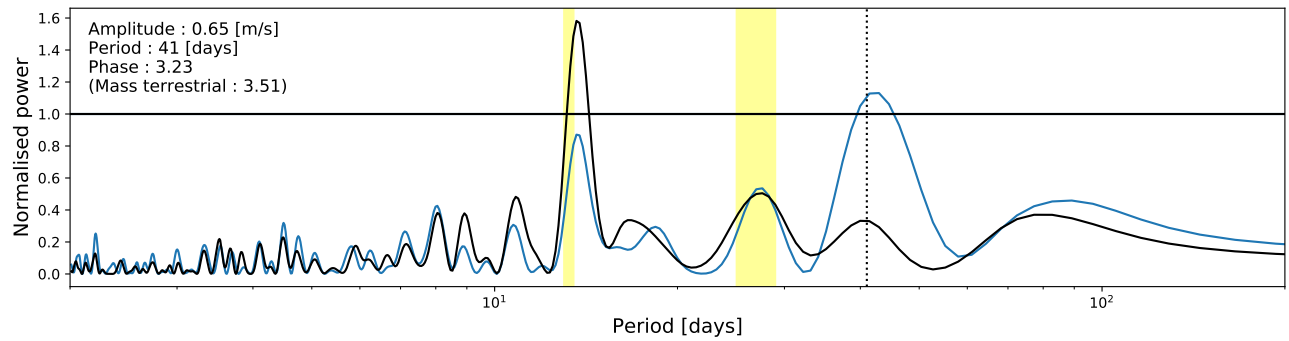

Fig. 57: Same as Fig.56 with a planetary signal : $K = 65$ cm/s, and $P = 41$ days (*dashed line*), $m = 3.5$ $M_\oplus$. The planet is not detected in the signal containing still activity.

We performed 2304 simulations with 24 different and evenly spaced periods and amplitudes (from 5 to 74 days and from 0.30 to 1.45 m/s respectively) and 4 different and evenly spaced phases. For a fixed amplitude and period, the detection rate is defined as the number of signals detected, among the 4 phases, divided by 4, thus the only possibilities are 0, 25, 50, 75 and 100 % of detection. Two distinct criteria are established based on the periodogram and on the MCMC results. For the former, we consider the planet as detected if the maximum power of the periodogram is no further than 10% of the injected period and higher than the 1% FAP level. For the latter, the planet is successfully detected if the three injected parameters (period, amplitude and phase) are inside the 2σ values around the median of the MCMC marginal likelihoods. We count as a non-detection if the planetary signal is compatible with a null signal (P or $K = 0$ inside the 2σ marginal likelihoods). The result of these simulations are displayed Fig.58.

With the periodogram criterion (left subplots), we note two «fingers» at $P \sim 14$ and $P \sim 28$ using all the lines (the former is clearly visible, while the second is shallower) which seems to show that even planets at 30 cm/s are always detected. Of course this is not the case and the signal found is in fact the activity one. This observation is supported by the formation of gap in the MCMC detection criterion (middle subplots) around the same periods. Even though the period found is the correct one (since the signal injected is close to the activity contribution) either the amplitude or the phase of the signal is wrong. The second gap is almost completely removed with our selection and only the first one is still present but less important (50% in the periodogram plot).

In the MCMC plots, some unexpected behaviours can be identified. For period around 8 days, our *uncorrelated* signal has a lower detection level than the one using all the lines. Also, for a fixed period, some low amplitudes are more easily detected than higher ones (for example around 50 days). In fact, this aspect is also in favour of our method. Sometimes, the MCMC simulations have difficulties to converge such that the likelihood is covering a huge zone of the parameter space : the parameters are less constrained (the standard deviation σ is higher). Using all the lines, the 1σ width of the amplitudes, periods and phases marginal distributions provided by the MCMC were in average 1.6, 1.3 and 1.3 times larger than the ones with our selection, hence the detections in our MCMC plot are samely 1.6, 1.3 and 1.3 times more difficult to achieve. To «restore the equity», we thus display in the last subplots the relative distance (difference) D_{rel} between the predicted parameters $\boldsymbol{p_{pred}}$ and injected ones $\boldsymbol{p}_{th}$:

$$D_{rel} = \frac{||\boldsymbol{p_{pred}} - \boldsymbol{p}_{th}||}{||\boldsymbol{p}_{th}||}$$

The median among the 4 phases has been taken. Let us precise that to account for the phase which is evolving in a closed dimension (modulo 2π), a mapping between the cylinder and the euclidean space has been performed :

$$\boldsymbol{p} = \begin{pmatrix} P/P_{th} \cdot \sin(\phi - \phi_{th}) \\ P/P_{th} \cdot \cos(\phi - \phi_{th}) \\ K/K_{th} \end{pmatrix}$$

As before, we detected the two fingers at $P \sim 14$ and $P \sim 28$ but also a new one around 50 days, which can be identified also in the periodogram Fig.56. This contribution could be explained by the high bumps visible around $BJD = 57580$ and $BJD = 57630$, two temporal windows where strong active regions have been observed (AR2565 - AR2567 and AR2585 - AR2587). The $P \sim 28$ and $P \sim 50$ contamination almost completely disappear with our lines selection. A new contamination not visible before is identifiable in our *uncorrelated* signal which is the second harmonic of the rotational period ($P_{rot}/3 \sim 9$ days).

In short, for the periodogram criterion, we can fix visually an amplitude limit of detection around 1 m/s against 60 cm/s using all the lines or our selection, respectively. For period longer than 60 days and small amplitude, less than two periods were available and the detection rate is lower. The relative distance (difference) limit of 20% in the parameter space is reached around 80 cm/s against 60 cm/s with our less contaminated signal but the better improvement is notable on the reduction of the frequency contamination. We end by indicating that a limit of $3M_\oplus$ can be attributed to our lines selection limit of detection, however this value cannot be transposed to the mass detection limit on other stars since the Sun has a very high SNR and is well sampled. This value may appear as disappointing by regards to some low-mass exoplanets ($M \sim 1M_\oplus$) discovered until now, however for a solar-type star, the minimum mass of a confirmed exoplanet discovered by RV is about $\sim 5M_\oplus$, for instance as around HD1461 (Díaz et al. 2016), 61 Vir (Vogt et al. 2010) or HD136352 (Udry et al. 2017). The current record[7] for a stellar mass $M > 0.9M_\odot$ is held by Kepler-406c with a semi-amplitude still highly uncertain $K = 1.0 \pm 0.6$ m/s. Moreover, these planets were discovered with sampling spread over a few years and have short periods ($P < 20$ days), in our case the detection were achieved with only 120 days of RV. Nevertheless, we have to remind that our simulations are only considering a subgroup of the Keplerian signals which are circular orbits for a star with an exceptional regular sampling and high SNR.

[7] https://exoplanetarchive.ipac.caltech.edu/index.html

We reproduced the same analysis using the maximum range in time of RV data in our possession, the 420 days after $BJD = 57330$ to avoid the strange offset observed. The *uncorrelated* selection was the union stars $\uplus$, the one giving the smallest rms for this particular timespan. The grid is made of 5224 simulations (Fig.59) of 23 and 4 evenly spaced values in amplitude and phase $(0.20 - 1.30$ m/s, $0 - 3\pi/2)$, respectively. The grid for the period is composed of 57 steps with three different "time-resolution". A high resolution ($\Delta P = 2$ days) was used for periods in range $2 - 34$, a middle resolution ($\Delta P = 3$ days) for $34 - 100$ and a low resolution ($\Delta P = 5$ days) for $100 - 200$ days. The improvement is less impressive than for the 120 days, in particular because the *uncorrelated* lines list seems to be less efficient in mitigating the stellar activity over such a long period. This is probably due to the fact that with the solar activity level of the Sun changing, not the same lines are affected by activity. Still, the planetary detection is pushed further with our selection, from 60 down to 40 cm/s. For long periods ($P > 140$ days), we note that the MCMC have more difficulties to converge toward the true signal, in particular since the uncertainties on the orbital parameters are underestimated. The focus on shorter period (Fig.60) allows to emphasise the mains frequency perturbations at 40, 28 and 14 days which are strongly mitigated with our *uncorrelated* selection.

As a final summary of the improvement, we represent in Fig.61 the difference of detection rate using our periodogram criterion between our line selection and the RV_m for the 120 and 420 days. All the red locations are detections improved, whereas the blue locations highlight a smaller number of detection. These blue locations were detections triggered by the activity in the RV_m which were producing however wrong amplitudes or phases in the MCMC. The only concerning degradation is two simulations in the 120 days data, at high amplitudes around $P \sim 60$ days. After a visual check, it appeared that the maximum power was situated at 12% of the injected periods. Because only two complete periods were available, the periodogram peaks were broad and in fact the true period was well located inside it, but slightly to far from the maximum.

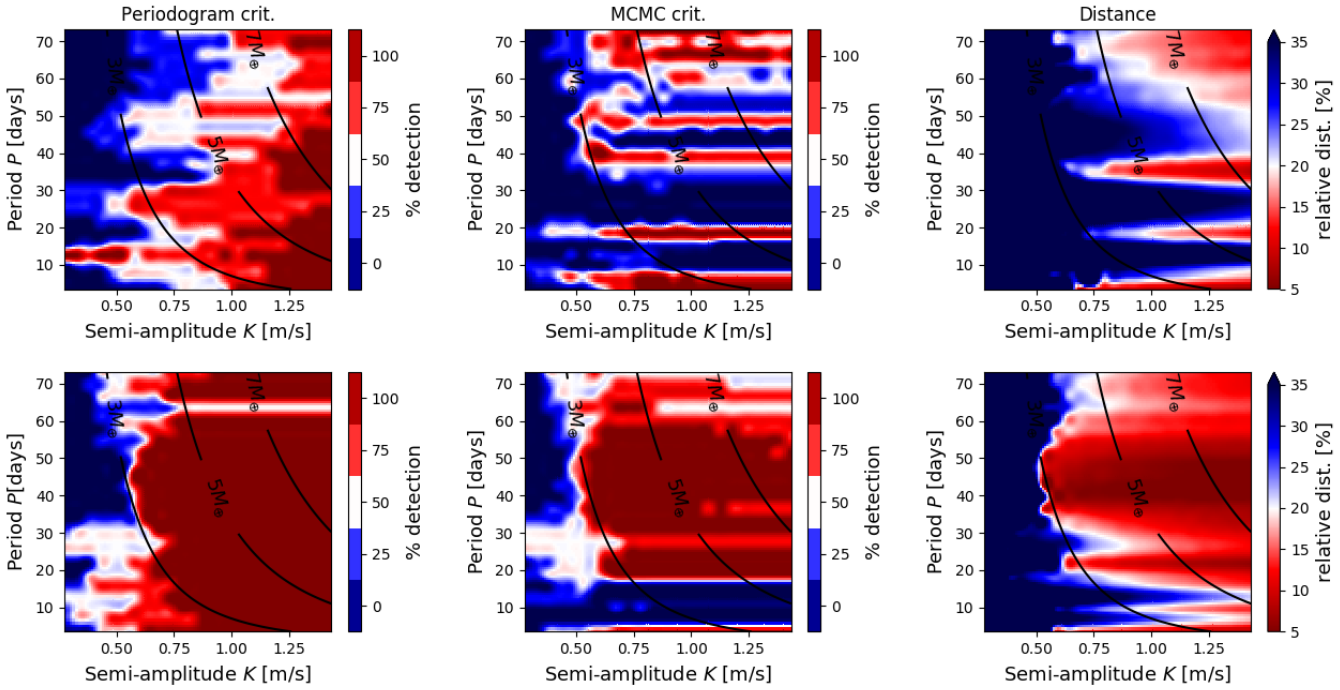

Fig. 58: Detection rate of the simulated planets injected (circular orbits) based on 2304 simulations made on a grid of 24, 24 and 4 evenly spaced values in period, amplitude and phase $(5 - 74$ days, $0.30 - 1.45$ m/s, $0 - 3\pi/2)$, respectively. Mass isocurves at 3, 5 and 7 $M_\oplus$ are indicated. **Top** : All the lines. **Bottom** : *Uncorrelated* lines. **Left** : Criterion based on periodogram. A planet is detected if the maximum power of the periodogram is less further than 10% of the true period and the peak is higher than the 1% p - value level. Two «fingers» are visible for period around 15 and 30 days using all the lines, because the signal detected is in fact the activity one. **Middle** : Detection rate based on MCMC. A planet is detected if the three injected parameters are in the 2σ marginal likelihoods. Two gaps are visible using all the lines at the period and half the stellar rotation period. With our *uncorrelated* selection, only the half period is still providing undetection. However, the marginal distributions are roughly 3 times smaller using our line selection such that our detections are 3 times more constrained. **Right** : Relative difference between the predicted parameters and the true values (median among the 4 phases). A maximum value of 35% of relative difference has been fixed in the color code. The predicted planetary parameters are closer to the true parameters injected when using RVs derived from with our lines selections. The main contaminations, still perturbing the detection, are the first ($P \sim 14$ days) and second harmonic ($P \sim 9$ days) of the rotational period. Let us precise that a small smoothing process has been applied on these subplots.

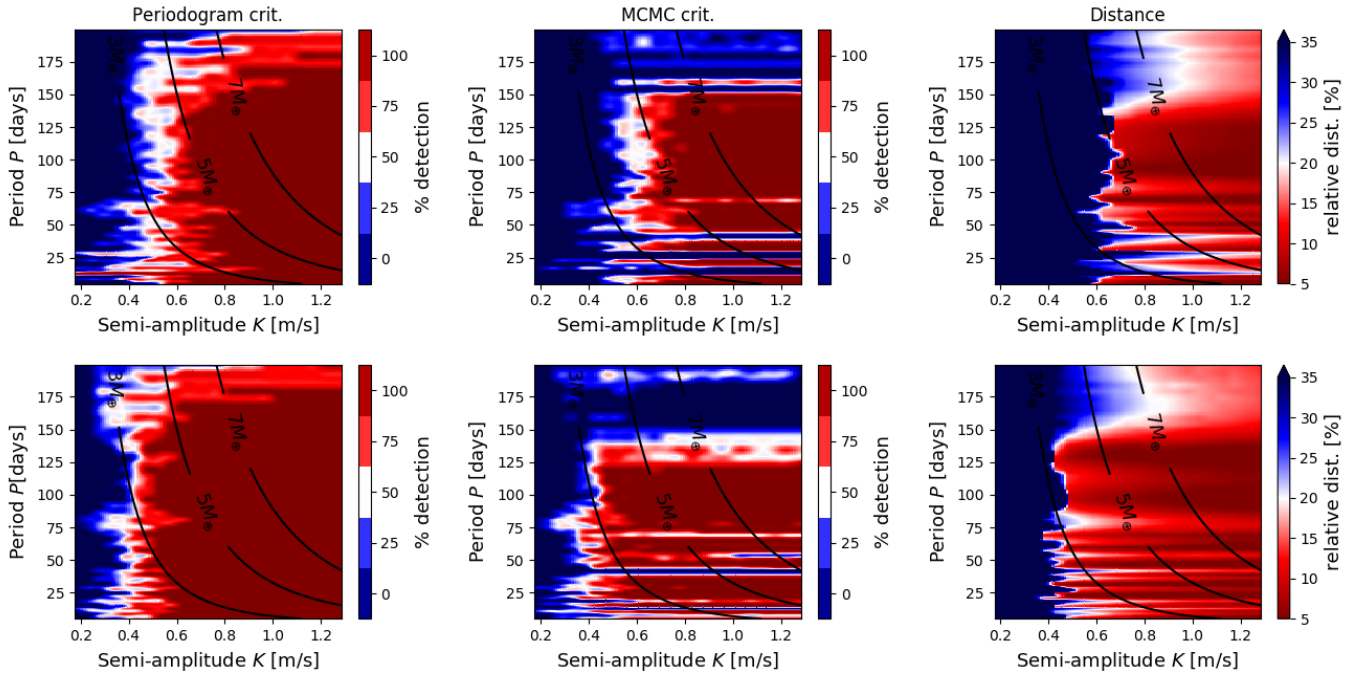

Fig. 59: Same as Fig.58 for the 420 days after $BJD = 57330$. The grid is made of 5224 simulations of 23 and 4 evenly spaced values in amplitude and phase $(0.20 - 1.30$ m/s, $0 - 3\pi/2)$. The grid for the period is composed of 57 steps with high resolution ($\Delta P = 2$) for $P < 34$ days, middle resolution ($\Delta P = 3$) for $34 < P < 100$ days and low resolution ($\Delta P = 5$) for $P > 100$ days. A focus on the short periods is displayed below (Fig.60). The improvement is less important than for the 120 days of RV data, coherent with the fact that the rms improvement were also less impressive. The planets are correctly recovered above 60 cm/s when all the lines are used, against 40 cm/s for our line selection. In the MCMC (2σ criterion), a brutal decrease is observed when less than 3 orbital periods are observed, regardless of the amplitude. We remark however that that when looking on the right plot to the relative differences between the predicted and the true orbital parameters, the recovered orbital parameters are still close of the injected ones ($< 25\%$), meaning that only their uncertainties, measured with the MCMC, are underestimated.

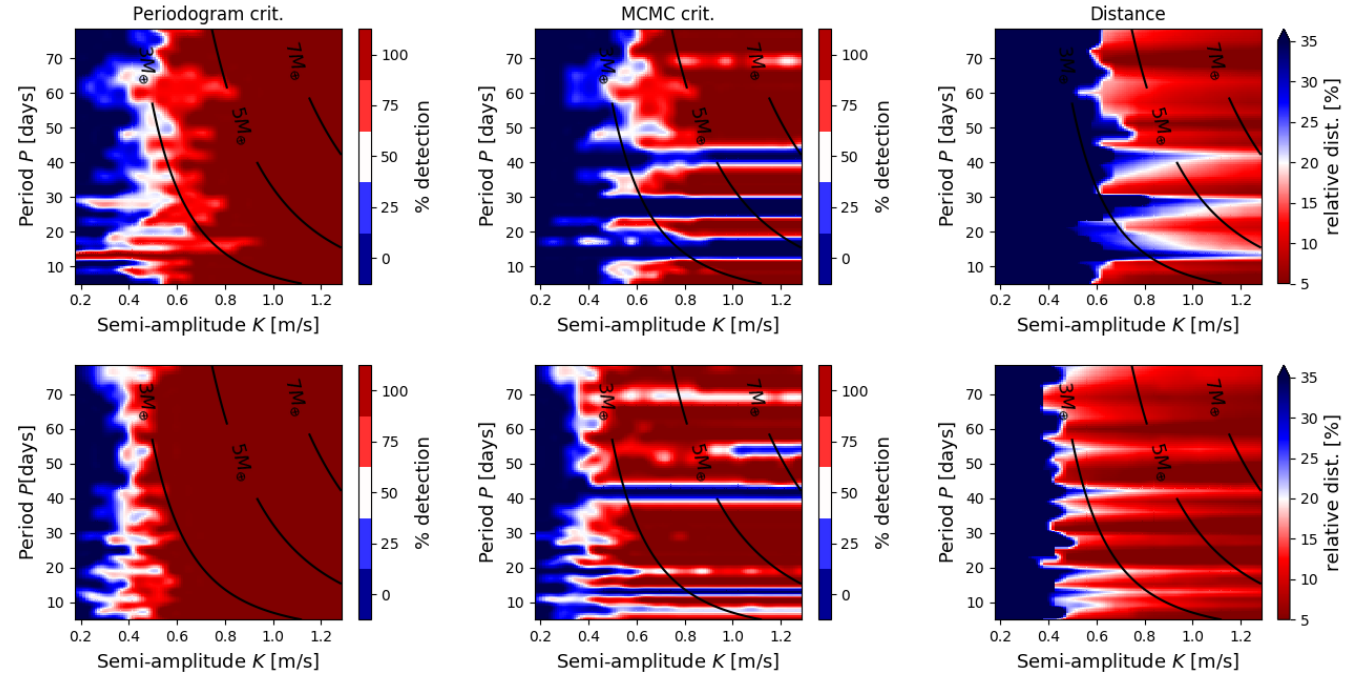

Fig. 60: Same as Fig.59 zoomed to provide a similar grid that in Fig.58. Three major contaminations are found around $P \sim 40$ days, $P \sim P_{rot} \sim 28$ days and $P \sim P_{rot}/2 \sim 14$ days in the RV_m, all these contributions are reduced when using our line selection. They are still visible in the MCMC criterion but disappear in the case of our line selection if instead of validating a signal when the orbital parameters are within 2σ of the injected values, we use a threshold of 3σ.

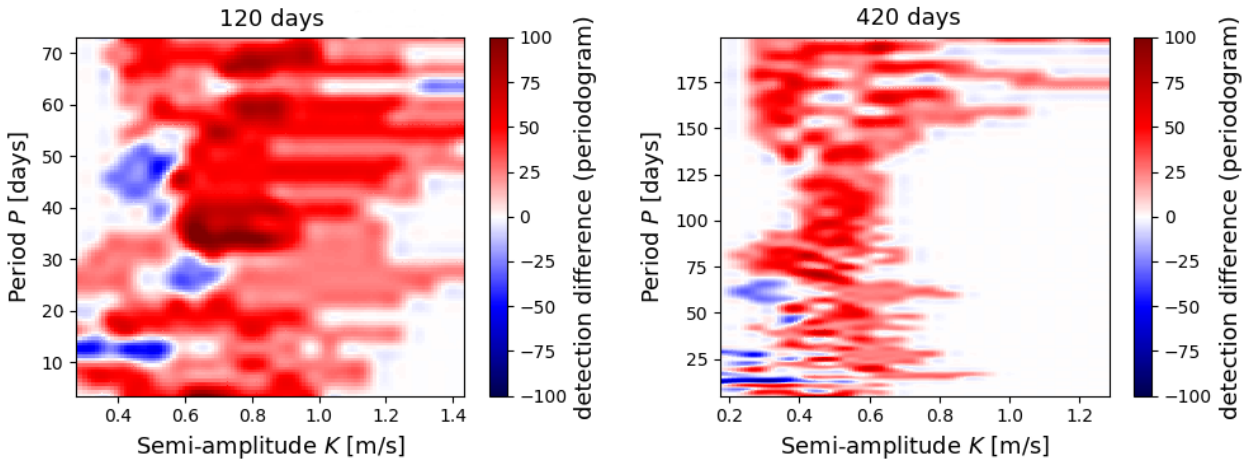

Fig. 61: Difference of the detection rate for the periodogram criterion in Fig.58 and Fig.59. White color means no improvement, red color emphasises improvements and blue color degradations. The degradations are in fact false detections induced by activity ($P \sim 14$, $P \sim 28$ and $P \sim 50$ days).

10. Conclusion

In this work, we analysed the spectral information coming from each individual line of α Cen B and the Sun with the purpose to better understand which lines are affected by activity or not. The first method based on flux variations has shown that a careful analysis has to be performed to avoid inherent bias present in ratio profile. Despite our attempts to correct for them, we do not find out evidences that this method was well suited to distinguish lines exhibiting high RV variation, probably due to a lack of precision (ratio profiles are necessarily noisier). A second method, based on machine learning algorithm and on the knowledges of the individual RV (RV_i), have on the opposite given signs that the atomic informations can be used partially to predict if a line will be or not affected by activity with an accuracy score of 70%.

We regret the complexity brought by the twin lines, i.e. the same spectral line that appears on two different orders in the spectrograph due to order overlapping, which rend the codes more sophisticated than necessary. Moreover, the presence of twins was a cause of concern at each step to understand how their treatment could impact the results : could they bias the machine learning predictions or the statistics ? What to do when the algorithm predict twins in different category ? A suggestion would be to average the information brought by the twin lines at the beginning of the analysis to avoid all these considerations.

Further improvements could be done on the optimisation of our *correlated* and *uncorrelated* lines selections. Indeed, we realised only too late that the optimisation can be simplified and faster dropping the RMS and SNR criterion (barely useless). Forcing the minimum slope of the *uncorrelated* selection to take the opposite value of the maximum slope would even bring the optimisation grid to a 2D level, against a 5 dimensional grid in our work. A new idea to form the groups may be to perform rolling correlations or a periodogram on each RV_i to disentangle the lines driven by activity.

The formation of two groups, lines affected versus lines not (or less) affected by activity, allows in one hand to efficiently reduce the rms of the 1-day binned RV signals down to $\sim$ 80 - 90 cm/s and to mitigate the long solar magnetic cycle down to 18 cm/s. At such level, supergranulation and instrumental limits are suspected to become relevant, but can be mitigate binning over several nights (rms $\sim$ 60 cm/s). In the other hand, measuring the RVs on the lines strongly affected by activity can boost the activity signal, which can help in characterising better stellar activity. Because such RVs are derived using less spectral lines, the white photon noise is increased. In the case where the photon noise is not dominant, like for Alpha Cen B and the Sun, stellar activity can be better characterised if its amplitude is increased, despite the fact that the white photon noise is larger. In particular, with the advent of the next-generation of spectrograph as ESPRESSO, the intrinsic line variation can be studied more easily since for this purpose the resolution is dominant on the SNR (Davis et al. 2017). We noted however that this binary classification is perhaps to restrictive to fully include all the shades of behaviours. The use of the score of certainty on each prediction could reject unreliable predictions, even if our exclusive union is precisely making indirectly the same job.

Good evidences are suggesting that the lines affected by activity change depending on the physical features of the active regions and the stellar properties. Thus, the swapping of selections between different stars is providing only a slight improvement and we find the same conclusion when swapping the selections optimised on different years (α Cen B year 2010 and 2011). The formation of an exclusive union provides a solution working quite well in general. We noted that the overlaps between the selections cannot be attributed to chance and that underlying property should be shared between the different lines selections.

The establishment of a synthetic activity signal which is mathematically deprived of planetary signals and its significant correlation with the $\log R'_{HK}$ index is a good indicator that our selection of line helps in mitigating stellar activity. We wonder if further investigations shouldn't be done on the SAS and its planetary-free property. For instance with a periodogram of the SAS and of our *uncorrelated* signal to use the SAS as a « filter », since a planet should form a peak in the latter, but not in the former (by lack of time we did not present results on this subject but planetary detection of 1.3 $M_{\oplus}$ was achieved).

The use of 10% of the spectral information when using the *uncorrelated* selection could be considered as an issue since it increases by a factor 3 the error bars on the RV signal compared to the CFF method. However, in our planetary simulations, we found that the orbital parameters recovered were roughly 1.4 times more precise, taking the advantage of a white photon noise compared to a red noise due to activity which perturb much more the planetary signal. In the case of the Sun, the uncorrelated selection of lines allows to detect planets with smaller RV semi-amplitudes. When looking at the detections limits in term of planetary masses, we observe that we are able to confidently detect planets down to three Earth masses, for periods shorter than 100 days, and even down to 2 Earth mass for periods shorter than 20 days.

Our feeling is that the Sun stays the most predominant target to understand the star activity since many counterparts (photometric observations, high resolution imaging) could be use to better characterise the actives regions, getting access to their latitude location, temperature and size which could feed the machine learning algorithms and help to categorised the individual line variation. Moreover, the wide knowledges of the solar system and solar cycle provide non-negligible advantage in the interpretation of the observed RV variations. Finally, this work is also the proof that constant and numerous observations on the same star are primordial if we want to mitigate the stellar effects. In the quest towards Earth-like planets, more efforts and programs should be attributed in the future to the development of deep investigations on less numerous stars in order to favor the quality of the understanding instead of the quantity of detections.

11. Acknowledgements

A special thanks to Xavier Dumusque who supervised me by day and by night when my motivation was unlimited, to Francesco Pepe for our discussions which have always grow the work in the right direction, to Jérémie Francfort and its cylindrical mapping solution and to the UNIGE and the attribution of the Excellence Master Fellowships which allowed me to study and work in comfortable conditions. My appreciations are also directed to my family and friends for their support and their interest (sometimes) in my work. This work has made use of the VALD database, operated at Uppsala University, the Institute of Astronomy RAS in Moscow, and the University of Vienna, and data coming from the ESO archive (Alpha Cen B) and from the HARPS-N solar telescope at the TNG in La Palma.

References

Anglada-Escudé, G., Amado, P. J., Barnes, J., et al. 2016, Nature, 536, 437
Baldi, P., Bauer, K., Eng, C., Sadowski, P., & Whiteson, D. 2016, Phys. Rev. D, 93, 094034
Beeck, B., Cameron, R. H., Reiners, A., & Schüssler, M. 2013a, A&A, 558, A48
Beeck, B., Cameron, R. H., Reiners, A., & Schüssler, M. 2013b, A&A, 558, A49
Berdyugina, S. V., Frutiger, C., & Solanki, S. K. 2001, in IAU Symposium, Vol. 203, Recent Insights into the Physics of the Sun and Heliosphere: Highlights from SOHO and Other Space Missions, ed. P. Brekke, B. Fleck, & J. B. Gurman, 254
Borgniet, S., Meunier, N., & Lagrange, A.-M. 2015, A&A, 581, A133
Cavallini, F., Ceppatelli, G., & Righini, A. 1985, A&A, 143, 116
Chapman, G. A., Cookson, A. M., Dobias, J. J., & Walton, S. R. 2001, ApJ, 555, 462
Davis, A. B., Cisewski, J., Dumusque, X., Fischer, D. A., & Ford, E. B. 2017, ApJ, 846, 59
Del Moro, D. 2004, A&A, 428, 1007
Del Moro, D., Berrilli, F., Duvall, Jr., T. L., & Kosovichev, A. G. 2004, Sol. Phys., 221, 23
Demory, B.-O., Ehrenreich, D., Queloz, D., et al. 2015, MNRAS, 450, 2043
Díaz, R. F., Ségransan, D., Udry, S., et al. 2016, A&A, 585, A134
Donati, J.-F., Yu, L., Moutou, C., et al. 2017, MNRAS, 465, 3343
Dumusque, X. 2012, PhD thesis, Observatory of Geneva <EMAIL>x.dumusque@unige.ch</EMAIL>
Dumusque, X. 2014, ApJ, 796, 133
Dumusque, X., Boisse, I., & Santos, N. C. 2014, ApJ, 796, 132
George, D. & Huerta, E. A. 2018, Physics Letters B, 778, 64
Gray, D. F. 2005, The Observation and Analysis of Stellar Photospheres
Hathaway, D. H. 2015, Living Reviews in Solar Physics, 12, 4
Haywood, R. D., Collier Cameron, A., Unruh, Y. C., et al. 2016, MNRAS, 457, 3637
Heiter, U., Barklem, P., Fossati, L., et al. 2008, in Journal of Physics Conference Series, Vol. 130, Journal of Physics Conference Series, 012011
Huang, X., Wang, H., Xu, L., et al. 2018, ApJ, 856, 7
Huélamo, N., Figueira, P., Bonfils, X., et al. 2008, A&A, 489, L9
Jones, D. E., Stenning, D. C., Ford, E. B., et al. 2017, ArXiv e-prints [arXiv:1711.01318]
Karttunen, H., Krüger, P., Oja, H., Poutanen, M., & Donner, K. J., eds. 2007, Fundamental Astronomy
Lecun, Y., Bengio, Y., & Hinton, G. 2015, Nature, 521, 436
Li, K. J., Feng, W., Liang, H. F., Zhan, L. S., & Gao, P. X. 2011, Annales Geophysicae, 29, 341
Maeder, A. 2009, Springer
Mayor, M., Pepe, F., Queloz, D., et al. 2003, The Messenger, 114, 20
Meunier, N., Desort, M., & Lagrange, A.-M. 2010, A&A, 512, A39
Meunier, N., Lagrange, A.-M., Mbemba Kabuiku, L., et al. 2017, A&A, 597, A52
Noyes, R. W., Hartmann, L. W., Baliunas, S. L., Duncan, D. K., & Vaughan, A. H. 1984, ApJ, 279, 763
Pepe, F., Mayor, M., Galland, F., et al. 2002, A&A, 388, 632
Pepe, F., Molaro, P., Cristiani, S., et al. 2014, Astronomische Nachrichten, 335, 8
Piskunov, N. E., Kupka, F., Ryabchikova, T. A., Weiss, W. W., & Jeffery, C. S. 1995, A&AS, 112, 525
Radick, R. R. 2001, in IAU Symposium, Vol. 203, Recent Insights into the Physics of the Sun and Heliosphere: Highlights from SOHO and Other Space Missions, ed. P. Brekke, B. Fleck, & J. B. Gurman, 78
Rajpaul, V., Aigrain, S., Osborne, M. A., Reece, S., & Roberts, S. 2015, MNRAS, 452, 2269
Reiners, A., Mrotzek, N., Lemke, U., Hinrichs, J., & Reinsch, K. 2016, A&A, 587, A65
Rieutord, M. & Rincon, F. 2010, Living Reviews in Solar Physics, 7, 2
Schrijver, C. J. & Zwaan, C. 2000, Solar and Stellar Magnetic Activity
Sousa, S. G., Santos, N. C., Mayor, M., et al. 2008, A&A, 487, 373
Thompson, A. P. G., Watson, C. A., de Mooij, E. J. W., & Jess, D. B. 2017, MNRAS, 468, L16
Toomre, J., Brun, A. S., De Rosa, M., Elliott, J. R., & Miesch, M. S. 2001, in IAU Symposium, Vol. 203, Recent Insights into the Physics of the Sun and Heliosphere: Highlights from SOHO and Other Space Missions, ed. P. Brekke, B. Fleck, & J. B. Gurman, 131
Udry, S., Dumusque, X., Lovis, C., et al. 2017, ArXiv e-prints [arXiv:1705.05153]
Usoskin, I. G., Solanki, S. K., & Kovaltsov, G. A. 2007, A&A, 471, 301
Vogt, S. S., Wittenmyer, R. A., Butler, R. P., et al. 2010, ApJ, 708, 1366
Wesselink, A. J. 1953, MNRAS, 113, 505
Wright, J. T., Marcy, G. W., Butler, R. P., & Vogt, S. S. 2004, ApJS, 152, 261

Appendix A: Unclassable lines

We only considered that lines could get either a RV signal similar to the average one, RV_m, or a flatter behaviour. One issue with the binary classification is that this limited vision does not account for diverging behaviours by respect to these two categories. In the figures below, we displayed some RV behaviours for the 2010 period of α Cen B. Except lines showing stranger behaviour, we remark lines anticorrelated (Fig.A.1) which is another interesting class, since such lines are necessarely related to the original active regions, even though they are moved in the opposite direction to the average signal. We tried to run the machine learning algorithms with the atomic parameters for a binary classification between *correlated* and *anticorrelated* lines, we also tried 3 classes *correlated*, *uncorrelated* and *anticorrelated* with even less success to understand what make special these *anticorrelated* lines.

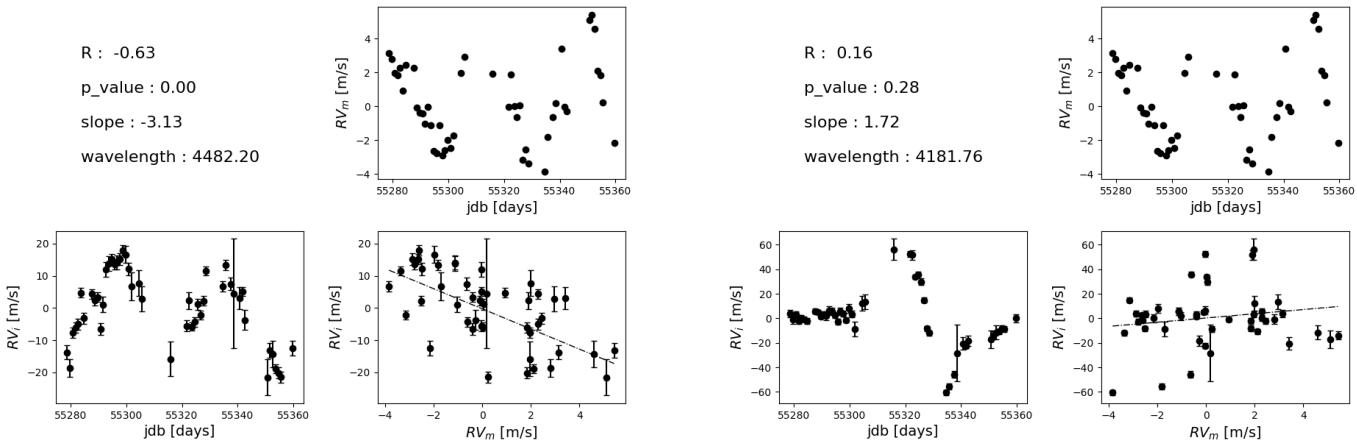

Fig. A.1: A few lines were detected as anticorrelated. This means that there is a relation between the RV of the line and stellar activity.

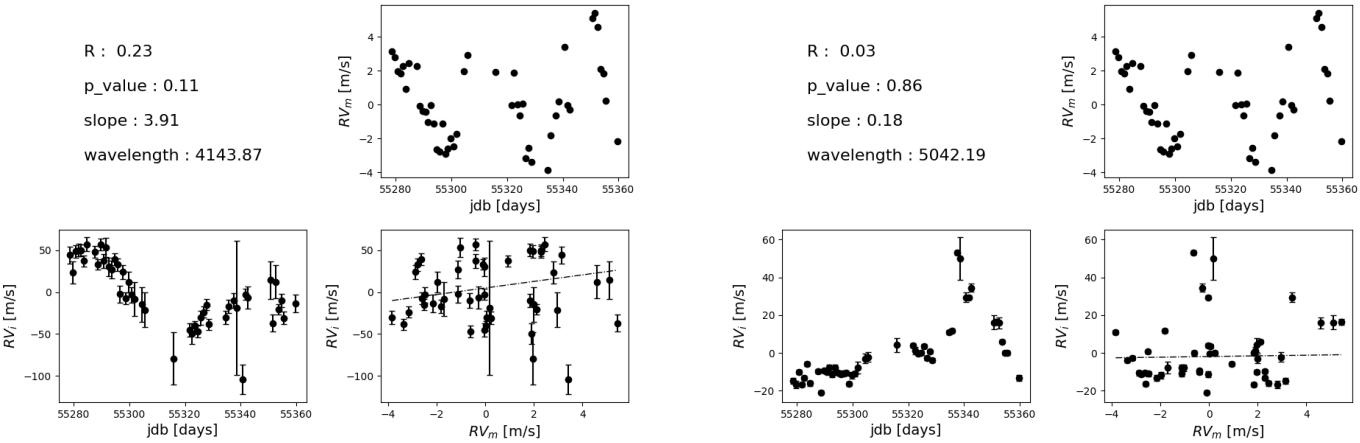

Fig. A.2: A line with a small dispersion but exhibiting a curious global trend. We note that the correlation slope is high, misclassifying this line as a correlated one.

Fig. A.3: Possible line *correlated* that show however a strong linear trend. Detrending the long term variation with a linear fit could bring back the signal to a similar shape than the RV_m.

Appendix B: Impact of the MAS on the Υ selection

There were uncertainties if the use of the 10 percent of most active night was a judicious choice in ratio profile since the spectra during the most active period are probably not similar due to variation in the position of active region on the stellar surface. In Fig.B.1, we draw the spectrum ratio using the 10 percent most active spectra, the most active night and the most active spectrum. The most active spectrum is definitely too noisy to be used,but the noise level is sufficiently low for the two others. In order to compare if the peaks produced by the most active night were different of those produced by the 10 percent most active nights, we built the confusion matrix Table.B.1. As highlighted by the barely diagonal aspect of the matrix, 80% of the peak are not changing of classes, this amount increase until 90% if we are only considering the *variation* and *no variation* classes. More *variation* peaks are found using the 10 percent night which is coherent considering that the noise is reduced when several spectra are averaged together. Note that this analysis has been performed before the discovery of the Thompson's bias.

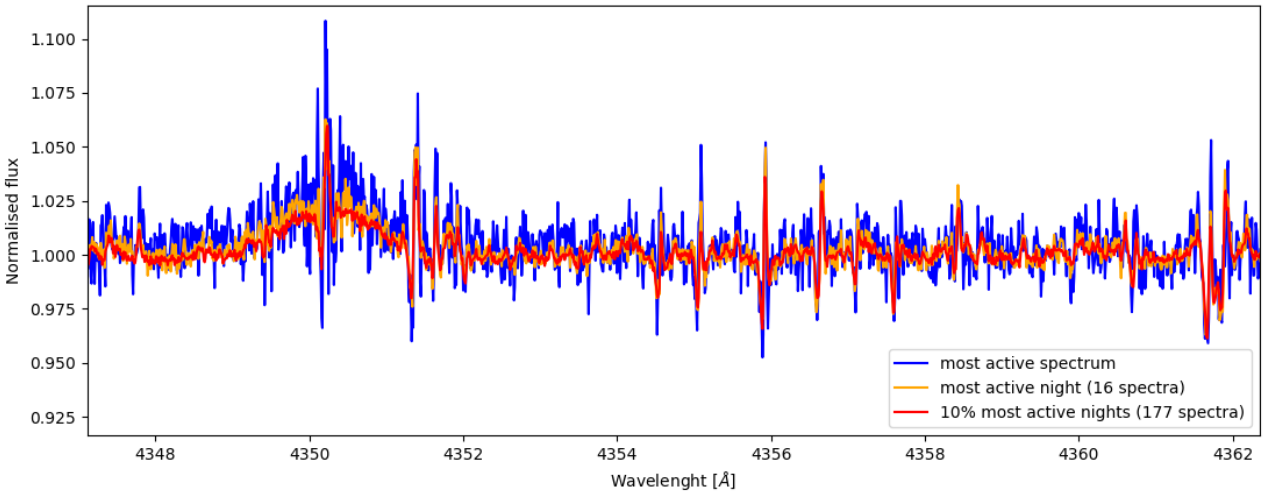

Fig. B.1: Spectrum ratio using different sample to form the *master active spectrum* (MAS). The sample used for this work, 10 percent of most active spectra (red curve), is compared to the most active night (orange curve) and the most active spectrum (blue curve).

Table B.1: Investigation on the consequence to use several or one night to form the MAS through a confusion matrix. The matrix is perfectly diagonal in the case of a perfect matching. The diagonal is highlighted in bold.

10% most act. \ most act.	No variation	Positive	Negative	Both	Σ_{raw}
No variation	**2966**	136	43	1	3146
Positive	260	**617**	13	62	952
Negative	141	25	**176**	28	370
Both	46	178	67	**370**	661
Σ_{col}	3413	956	299	461	5129

Appendix C: Examples of confusion matrix

Despite the strong advantages bring by the machine learning algorithms, the user should still keep in mind that these tools are only powerful if the user behind keeps an attentive eye on the results, for instance checking carefully the confusion matrix of the test session. Below are displayed two commons bad simulations that can occur in a case of a binary classification between A and B classes.

Appendix C.1: Confusion matrix of a random selection

The first bad scenario would be that our algorithm doesn't find any way to distinguish our classes. This issue would provide a *random classification*. Assuming a bag containing N letters. With N_A and $N_B = N - N_A$ letters. For a random choice algorithm the value of the recall, the precision and the F^1 test will converge to :

$$\Pi_j^{random} \simeq \frac{N_j}{N} \qquad R_j^{random} \simeq 50\% \qquad F_j^{random} \simeq \frac{2 \cdot \frac{N_j}{N}}{1 + 2 \cdot \frac{N_j}{N}}$$

Whereas the accuracy A^{random} and the error rate E^{random} will both $\simeq 50\%$. Let's remark than even though all parameters are not equal to one half, when the mean is taken between the two classes : "A" and "B", this time all the statistics converge to 50% as expected for random choices between two classes.

Table C.1: Example of a confusion matrix of a random choice algorithm. The bag of the simulation is filled with N letters, N_A are A and $N - N_A$ are B. After the prediction of the algorithm, both the accuracy A and the error rate E are at $\sim 50\%$.

True \ Predicted	A	B	Σ_{row}	R
A	$\frac{N_A}{2}$	$\frac{N_A}{2}$	N_A	0.50
B	$\frac{N-N_A}{2}$	$\frac{N-N_A}{2}$	$N - N_A$	0.50
Σ_{col}	$\frac{N}{2}$	$\frac{N}{2}$	N	
Π	$\frac{N_A}{N}$	$1 - \frac{N_A}{N}$		$\frac{2 \cdot \frac{N_A}{N}}{1+2 \cdot \frac{N_A}{N}}$ & $\frac{2 \cdot (1 - \frac{N_A}{N})}{1+2 \cdot (1 - \frac{N_A}{N})}$

Appendix C.2: Confusion matrix of a biased « overpredicted » selection

The second dangerous scenario would be that our algorithm decides to always predict the same class. Assuming a bag containing N letters. With N_A and $N_B = N - N_A$ letters. For a biased algorithm (without loss of generality here A), the value of the recall, the precision and the F^1 test are written in the confusion matrix shown in Table.C.2. The accuracy A^{biased} will be equals to $\frac{N_A}{N}$ and the error rate $E^{biased} \simeq 1 - \frac{N_A}{N}$. We can wonder why an algorithm would make such a strange choice, but in fact it can occur if during the training, the two groups have a large difference in size. If for example, the group B only represents 5% of the bag, the algorithm will naturally make most (even all) its predictions as A letters, even if during the test session there is the same amount of A and B letters. That is why it is important to control during the training session that the two groups are not disproportionate one compared to the other. Fortunately, such bias are identifiable because they produce a kind of « gap » between the two recalls.

Table C.2: Example of a confusion matrix of a biased algorithm. The bag of the simulation is filled with N letters, N_A are A and $N - N_A$ are B. After the prediction of the algorithm, the accuracy A^{biased} will be equals to $\frac{N_A}{N}$ and the error rate $E^{biased} \simeq 1 - \frac{N_A}{N}$. Such bias can be identified by the formation of a « gap » between the two recalls.

True \ Predicted	A	B	Σ_{row}	R
A	N_A	0	N_A	1
B	$N - N_A$	0	$N - N_A$	0
Σ_{col}	N	0	N	
Π	$\frac{N_A}{N}$	undef		$\frac{2 \cdot \frac{N_A}{N}}{1+\frac{N_A}{N}}$ & undef

Appendix D: Machine learning with only physical parameters

We display below the resulting classification of relevant parameters according to the logistic machine learning algorithm and XGB after suppressing the observationnal parameters such as the $\mathcal{R}$ or the FWHM_RHK correlation slope. Both algorithms predict that the main physical parameters are the line depth and wavelength of the transition.

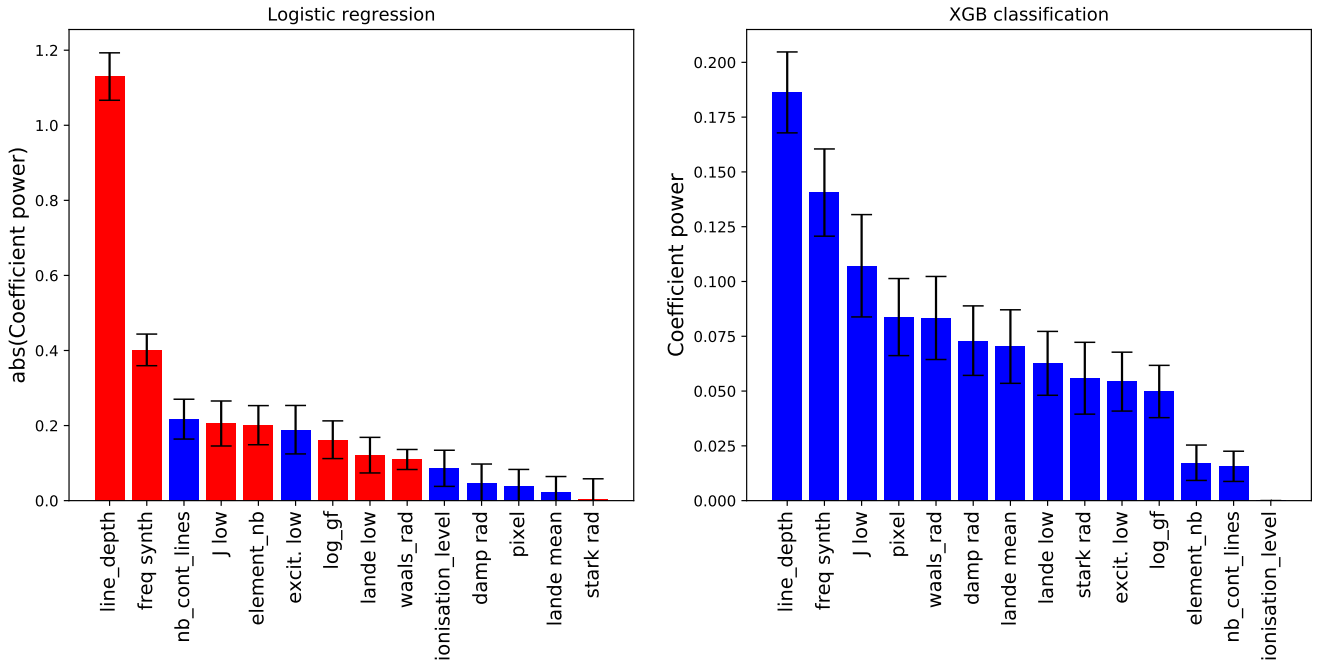

Fig. D.1: Same graphic as Fig.30 suppressing two observational parameters which could be missing for others stars. The logistic algorithm is still dominated by the line depth whereas the XGB seems to account for more parameters in its decision (its score is by the way better than the logistic one).

Appendix E: Counterpart in SOHO and SDO imaging

To understand if the offset visible in the 1.5 years RV data of the Sun was instrumental or linked with activity we looked the images provided by the extreme ultraviolet imaging instrument (EIT) on board of SOHO. The images are taken at 17.1 nm. Before the offset, 14^{th} October 2015, only one active region was visible and the 2^{nd} was just emerging of the limbs. The high RV observed around the 20^{th} October 2015 is compatible with the unusual number of emission observed in UV. The jump produced the 19^{th} October 2015 is coherent with the new regions appearing on the left side (the actives regions 3 and 4) and the fact that the RV were maximum the 22^{th} October 2015 is the time delay for the active regions to reach the center. The active regions are also visible in HMI magnetogram from SDO. Our trouble was born initially because this offset was not observed in the $\log R'_{HK}$, after private communication it appears that the HARPS-N was warmed up precisely at this time, 13^{th} October 2015, which had coincidentally for effect to smooth the $\log R'_{HK}$ instead of providing a discontinuity.

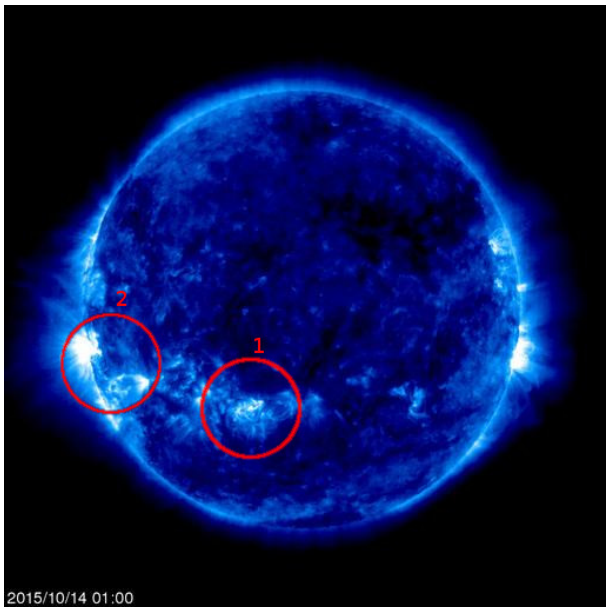

Fig. E.1: SOHO imaging by the EIT 171 instrument the 14^{th} October 2015, almost no activity is visible. (*Crédits : SOHO imaging*)

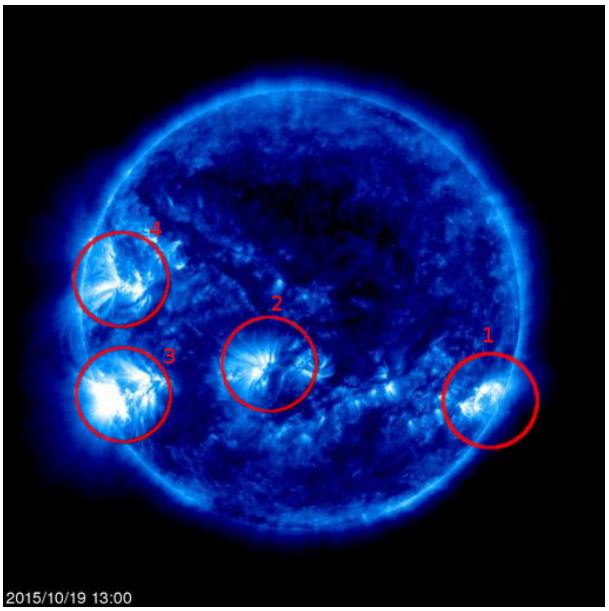

Fig. E.2: Same as Fig.E.1, the 19^{th} October 2015. At least 4 emission regions are visible in UV and many spots were also detected in continuum imaging. (*Crédits : SOHO imaging*)

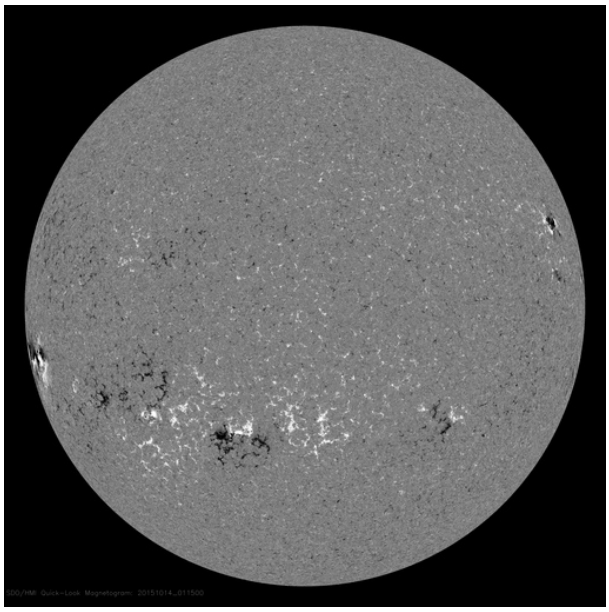

Fig. E.3: Same as Fig.E.1 for SDO HMI magnetogram. (*Crédits : SDO imaging*)

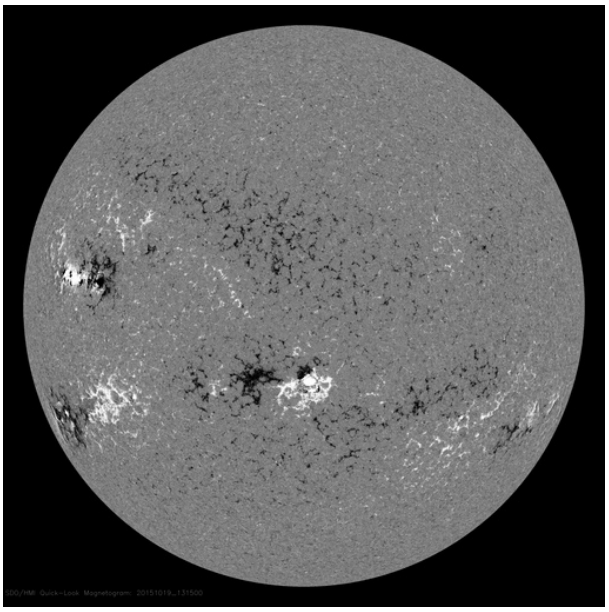

Fig. E.4: Same as Fig.E.2 for SDO HMI magnetogram. (*Crédits : SDO imaging*)

Appendix F: Matrix plot seaborn

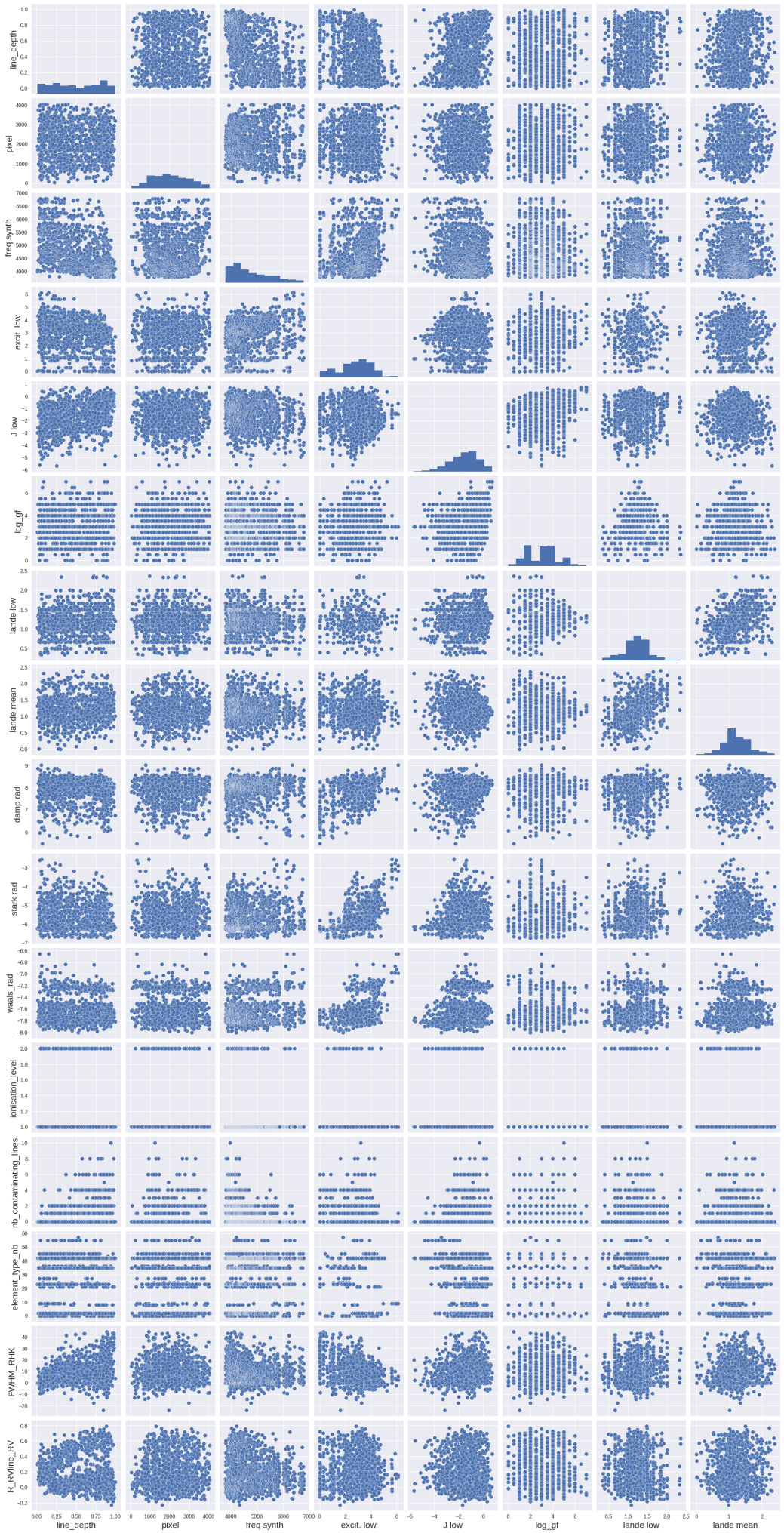

Fig. 1: Parameters pair plot (made by the seaborn library) for the 16 parameters used in the machine learning analysis for α Cen B in 2010. We visually checked that the parameters were not correlated ones with the others. From top to bottom (or left to right) the parameters are : line depth, pixel, wavelength, excit. low, J low, log(gf), lande low, lande mean, damp rad, stark rad, waals rad, ionisation level, nb cont. lines, element, FWHM_RHK, $\mathcal{R}$

al velocities